Essays in Optimizing Social Policy for Different Populations: Education, Targeting, and Impact Evaluation

Essays in Optimizing Social Policy for Different Populations: Education, Targeting, and Impact Evaluation

A dissertation presented

by

Sara B. Nadel

to

The Department of Public Policy

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Public Policy

Harvard University

Cambridge, Massachusetts

December 2015

Dissertation Advisor:
Professor Lant Pritchett

Author:
Sara B. Nadel

Essays in Optimizing Social Policy for Different Populations: Education, Targeting, and Impact Evaluation

Abstract

In the first chapter of this dissertation, I look at the relationship between preference sets among students in similar majors, compared with different majors, in Peru. I find that students within majors share preference sets that differ from students in other majors. I further find that students from households without a formal labor market participant have made decisions that are more consistent with predicted professional opportunities compared with students with a formal labor market participant. These differences are systematic and not related to the general industrialization level of the city where the student lives. This research suggests that the difference between students and workers from households with formal labor-market familiarity and those from households without formal labor-market familiarity are not accidental or due to lack of familiarity.

In the second chapter, I evaluate whether proxy-means testing as a method of targeting for Mexico's Conditional Cash Transfer program caused spending distortions among (potential) recipients. The income and wealth effect of participating in *Progresa* complicate a simple comparison of members of the control and treatment group in the acquisition of assets. To resolve this, I look at reduced asset acquisition just above the cutoff point. Because an imperfect implementation of the eligibility evaluation may have reduced treatment villagers perceived benefit of distorting, I also look for evidence of increased spending in non-assets and of increasing the number of eligible-aged children in the home to increase the size of the transfer. I do not find evidence of lack of investment in assets along the eligibility cutoff, but I do find evidence of increased spending as a percentage of income on items not

included in the PMT, as well as evidence of increases in eligible-aged children among the poorest families in treatment villages.

In the final chapter, which is joint with Lant Pritchett, we propose that many development programs, projects and policies are characterized by a *high dimensional design space* with a *rugged fitness function* over that space. In nearly any project/program/policy there are many design elements, and each design element has a number of possible choices, and the combination produces a high dimensionality design space. If different program designs produce large changes to outcomes/impact, this implies that the "fitness function" or "response surface," the mapping from program design to outcomes/impact, is rugged. We motivate this investigation using as an example a skill-set signaling program for new entrants to the labor market in Peru. We present a simulation model which compares two alternative learning strategies: "crawling the design space" (CDS) and a standard randomized control trial (RCT) approach. In this artificial world, we demonstrate that with even modest dimensionality of the design space and even modest degrees of ruggedness, the CDS learning strategy substantially outperforms the RCT learning strategy. Moreover, we show that the greater the ruggedness of the fitness function, the higher the variance of the RCT results relative to CDS and hence the lower the reliability of RCT results even with "external validity" across contexts. We suggest that RCT results to date are consistent with a world in which social programs exist in a high dimensional design space with rugged fitness functions and hence in which the standard RCT approach has limited direct practical application.

Contents

List of Tables

List of Figures

Acknowledgments

It takes a village to complete a dissertation, and my village was particularly generous and extensive.

This paper would not have been written without the support of Lant Pritchett, whose insights into the value of the research and whose mentoring about how to actually complete it were invaluable. Matt Andrews and Daniel Levy supplemented this guidance with terrific feedback and support. I also benefited from advice at different points in this process from Josh Goodman, Rema Hanna, Dean Karlan, Michael Kremer, Amanda Pallais, Rohini Pande, Juan Saavedra, and Richard Zeckhauser, the participants in the Harvard Economics / Kennedy School Development Lunch, the and staff at Innovations for Poverty Action, Peru.

I am grateful to The Hauser Center for Nonprofit Organizations and the Women in Public Policy Center Adrienne Hall Fellowship for financial support.

Vanessa Carella, Vilma Grima Estrada, Alexandra Goldenberg, Mayra Narvaez, and Karen Ramos provided on-the-ground support for the field research in Peru and pushed the research forward through the pivots and reversal documented in this research. Tom Barry has been a crucial mentor to me throughout my professional career.

Angela Fonseca, Jeff Friedman, Mahnaz Islam, Chris Robert, Anitha Sivasankaran and Ana Tribin participated in regular research groups with me at various points in the process. Caitlin Eicher, Maggie McConnell, Ariel Stern, Tisa Sherry and Liz Walker formed the broader academic brain trust from whom I learned how to organize myself in order to write a dissertation. Diane and Lant Pritchett were supportive hosts in their beautiful mountain residence during a several-day slog through my joint paper with Lant.

My parents, Susan Bryson and Laurence Nadel, and my siblings, Gabe Nadel and Sophia Skaar, made up a family in my childhood where knowledge and investigation is highly valued, particularly when it takes you to far flung corners of the globe. My grandparents, James and Jane Bryson and Herbert and Millicent Nadel, set the foundation for these family values. Marshall Cox complements and improves them.

All errors are my own.

Chapter 1

Preferences, Career Choice, and Contact with the Formal Labor Market

1.1 Introduction

Choosing a major, and hence a career, is an exercise in long-term decision-making under limited information. Information is limited about one's own skillsets and which professional activities one enjoys, what skillsets and preferences are required for each position, and what the demand for each type of professional will be over the career arc of the individual's career.

This paper examines how students with less information about the formal labor market make different career decisions than those with more information about the formal labor market. I look at the preference sets of university students in Peru by choice of major. I find that students in different majors have different preference sets. I also find that, within majors, students from households where nobody has a formal job have different preferences than students from households where there is a participant in the formal labor market. I find cases in which these differences are systematic, as opposed to random, suggesting that individuals from households without a participant in the formal labor market are either optimizing their choice of major over different preferences or over systematically different

(mis)information. I do not comment on the potential differences between the experience of *studying* for a particular career and actually *practicing* that career upon completion of studies, although it is very likely that those differences play a role in the completion of a given major.

The way that poor, upwardly mobile populations fit into the labor market is an important topic. If they are choosing careers they will not enjoy because of poor information or expectations, they are at a higher risk of dropping out of the workplace. On a macro scale, this phenomenon as a hazard of fast-industrializing nations could help explain the reducing returns to education (Pritchett, 2001). If first-generation university students arrive at university with a different cultural orientation, but the university experience facilitates a preference-set convergence, school may play a valuable non-educational role in preparing students for the formal labor market. Alternatively, formal labor markets could adjust to make jobs more agreeable to the new entrants into the labor market.[1]

My contribution to this field is to demonstrate that preferences play a role in career choice, and that familiarity with the formal labor market may impact how students choose their career.

The rest of the paper continues as follows: In Section 1.2, I discuss literature about the relationship between preferences and career choice and success. In Section 1.3, I present the data used in this study. I review the economic environment in Peru in 2013 and 2014, the partner organization, and the preference measurements. In Section 1.4, I review the data and implementation. Section 1.5 presents my hypothesis and Section 1.6 summarizes results. Section 1.7 concludes.

1.2 Relevant Literature

Research about career sorting and learning, and optimizing over one's own skillset is frequently based on the Roy Model, which proposes a labor-matching outcome in which

[1]Many firms in Peru that hire first-generation formal labor market participants (attempt to) do this.

workers take the jobs where they have a comparative advantage (Roy, 1951).

This paper draws on literature about the role of preferences, values, and career choice. Satterwhite (Satterwhite et al., 2009) and Bradley-Geist and Landis (Bradley-Geist and Landis, 2012) present the Attraction-Selection-Attrition (ASA) model for occupations (ASA had previously been used for organizations), whereby the people within an occupation tend to be homogenous. Similar people are attracted to similar occupations, those with the power to select (hire) choose people who are similar to those already in the occupation. People who do not fit in eventually leave. This study proposes that people with little knowledge of different vocations are less effected by the *attraction* part of the model; where the university accepts all students, *selection* has no role in ASA; and so, students who mis-sort into a vocation that does not align with their preferences and values are more likely to *attrit*. Holland (Holland, 1985) demonstrates that people with the same vocation are similar in personality, establishing personality and preferences as a legitimate basis for "similarity" in the ASA model.

Several structural models have contributed to an understanding of how students choose majors or careers in an uncertain environment. Kinsler and Pavan develop a structural model to evaluate the wage premium of working in a job related to one's college major. They find that the premium was 30% for science majors. However, when they control for skills at the start of their career and those acquired over the course of study, they find that this premium disappears, suggesting that choice of major is a first step in the ASA process. Their findings reduce the risk of choosing a poor-fit major in a labor market with flexibility to change occupations after graduation as one learns about one's skillsets and preferences, but not in a more rigid one such as Peru. Their model about career choice under uncertainty about the labor market and one's own skills informs the model presented here (Kinsler and Pavan, 2012).

Sullivan develops a structural model to evaluate the gains from dynamic matching in the workplace: employees moving jobs as they gain new information about themselves (Sullivan, 2010). He finds the gains to be as high as 31%, suggesting that a rigid labor market that

discouraged ongoing matching would experience less improvement in productivity.

Previous studies have looked at how students with less information about careers choose their careers or majors. Simpson and Simpson (Simpson and Simpson, 1960) interview college students about their values (rank preferences), and their career choice, and find that students tend to choose careers where the majority of other students have similar values as they do, and that the mean value profile differs by career. Phillips (Phillips, 1968) finds that boys with weaker sources of information about their chosen vocation tend to have preferences less congruent with the other boys who had chosen that vocation. Pallais evaluates the effect of increasing the number of schools to which college applicants can send their ACT scores for free. She finds that students applied to a broader range of schools, with the outcome that they attended more competitive schools, increasing long-run earning potential (Pallais, 2013). Lack of information about the college-application process compromises earnings potential. Van der Klaauw highlights the role of limited information and uncertainty among individuals in choosing profesional paths by comparing the expectations of graduates from a teaching school with a structural model (developed post-facto) which is stronger at predicting graduates' professional outcomes than they themselves were (Van der Klaauw, 2012).

Jensen demonstrates that students' perceptions about the benefits of higher education are skewed by the communities in which they live because their communities are more income-homogenous than the population in general and are made up of either people who went to college and had average outcomes and people who didn't go to college and had above-average outcomes, or the reverse (Jensen, 2010) .

In the absence of large amounts of information about the labor market, job-seekers rely on other sources of information, with varied results. A large body of research examines the role of internal references in securing a job. One argument suggests that people who secure jobs through recommendations are higher performers because the recommender is relying on private information about the quality of the job fit (Breaugh and Mann, 1984). Other research demonstrates a matching mechanism whereby recommended candidates are lower

quality because the recommender is exploiting the benefits of making a recommendaton and the social capital earned from doing so, as in the Ghanian army (Fafchamps and Moradi, 2009). Some recommended candidates are lower-quality workers, but are beneficial to their employers in that they churn less because they don't have other outside options (Loury, 2006).

Research about the development and importance of non-cognitive and cognitive skill finds that these characteristics are important determinants in professional success. Heckman et. al. suggest the importance of non-cognitive skills in professional success (Heckman, Stixrud and Urzua, 2006). Heckman finds that both cognitive and non-cognitive skill development occur at a very early age, and are heavily influenced by family environments (Heckman, 2006). He suggests that children from lower-income households have, on average, lower cognitive and non-cognitive skills than wealthier children. If students from families without a participant in the formal labor force are poorer than other students (not necessarily the case in industrializing economies), then they may begin their college education at a cognitive and non-cognitive skill disadvantage, regardless of the congruency between their values and career choice.

Macro characteristics have also been used to explain lower productivity in developing countries. Tybout finds that manufacturing productivity in developing countries tends to be lower than in richer countries (Tybout, 2000). He proposes that this is due to a a lack of competition, smaller size, and poorly functional financial markets. Bloom et al. add to this finding, suggesting that firms in developing countries are badly managed and the highest management levels are typically filled by family members. Since family members churn less, and churn is an important way for firms to learn best practices, learning and improvement is limited (Bloom et al., 2010). These characteristics certainly also play a role in lower productivity in developing countries.

1.3 Background

1.3.1 The Peru Higher Education System

Peru offers many types of tertiary education, all of which are viewed as important tools to secure a better job. The most common of these are university degrees (typically 10 semesters), and technical degrees (typically 6 semesters). The government is heavily involved in the development of the curriculum: it identifies the list of acceptable majors, and requires that all university degrees include at least one semester of an internship within the given major.

Universities are the more prestigious degree. Students who complete the curriculum receive *Bachilleratos*. Students frequently go on to complete a *Licenciatura*, which requires either a thesis and/or an additional high-level course, and is typically pursued over weekends and in the evenings once the student is working full time.

The university system includes public national universities, which are run by the government, and private universities, which are for-profit entities. The national universities are less expensive than private universities and are well respected in the quality of their education. However, they are more susceptible to closure due to strikes, and thus students sometimes report choosing a private university if they are eager to finish their degree within a set period of time.[2]

The number of private, for-profit universities has been growing in Peru. Many private universities, including the partner in this study, have locations in several cities and continue to add locations. They typically cost more than national universities and, with a few exceptions, are known to have a lower bar for acceptance.

University acceptance is entirely test-based. Once a semester, the university offers a test, for which there is a fee (at our partner university, this fee is S/.10, or about $3.70).[3] Students apply for a specific major. Unlike the US, where the SAT or ACT is accepted at most universities, in Peru, every university has its own test, and students who wish to apply

[2]Informal qualititative interviews with university students.

[3]Based on July 2014 exchange rate.

to several universities must pay for and take several tests. Popular universities including the National Universities have limited slots available for each major, and accept as many students as slots available, entirely based on score. More popular majors, such as Industrial Engineering, have a higher score cutoff than less popular majors, such as Education. Private universities do not publish acceptance rates, but it is believed that many private universities accept all applicants above some minimum score, so that the popularity of the major does not contribute to the probability of acceptance.[4]

Demand for major is reflected in price. Prices are higher for more popular majors, for majors with higher earning potential, and for majors with a higher cost of supplies. At the partner university in this study, Accounting costs S/. 1,750 ($645) per semester, Organizational Psychology costs S/.1,900 ($703) per semester, and Early Education costs S/.1,250 ($462) per semester (in July, 2014).

The choice of major is important. Students cannot switch majors without re-applying to the university, and courses from one major can rarely be applied to courses from another. Furthermore, it is extremely difficult to secure a professional job outside of ones general area of study, if not one's own major. Even the choice of internship is important: if a psychology student does their internship in clinical practice and decides afterwards to pursue a job in human resources, he or she would need to do a second internship in order to secure the sought-after position. As a result, students make long-term decisions about their professional future immediately out of high school, when they choose their major and apply to university.

1.3.2 The Peru Labor Market

The job market in Peru, particularly for jobs requiring some higher education, has grown in the past ten years. Between 2001 and 2012, the number of formal jobs grew by 31.0%; but the number of formal jobs for people with at least some higher education has grown 98.7% (Ministerio de Trabajo del Peru Dataset, 2012).

[4]One higher-education institution has confirmed this with me off the record.

This growth is particularly high in some of the cities where the UCV has locations. Table 1.1 and Table 1.2 show the growth in all jobs and jobs requiring some tertiary education in the cities included in this study, respectively.

Table 1.1: *Labor Market Growth, 2001-2012*

Location	**Percent growth, Number Formal Workers**	
	w/at least some Higher Education	All Workers
Chimbote	116.0%	25.4%
Lima (Lima Norte & San Juan de Lurigancho)	81.7%	38.3%
All Peru	**98.7%**	**31.0%**
Piura	113.0%	28.3%
Tarapoto	141.6%	36.2%
Trujillo	125.1%	40.3%

Table 1.2: *Perc. Formal Workers with some Higher Education*

Location	**Makeup of jobs & labor force: Higher Education**	
	Some tertiary education, 2001	Some tertiary education 2012
Chimbote	15.9%	27.4%
Lima (Lima Norte & San Juan de Lurigancho)	32.2%	42.3%
All Peru	**20.9%**	**31.7%**
Piura	15.6%	25.9%
Tarapoto	13.3%	23.6%
Trujillo	17.7%	28.4%

The increasing demand for workers with some higher education may be driven not only by an increase in the demand for skilled labor, but some level of inflation in worker requirements due to an increase in the supply of skilled labor. As more tertiary schools have cropped up, so has the portion of the worker-age population with some higher education increased. In this paper, I do not attempt to separate out these dueling effects, nor do I discuss the quality of the new higher education offerings.

1.3.3 Partner University

This paper uses data from an implementation of the Farolito Test, a skills and preference-assessment test, at a for-profit university in Peru. It was founded in the early 1990s and has opened locations in 11 more cities since its inception. The university growth strategy is to establish in a new location and slowly add majors. In this rollout strategy, it is similar to other expanding private universities in Peru. As a result, not every location offers each major, and several majors are offered for incoming students, but do not yet have graduating students.

As a for-profit university, the partner university is more expensive than its national (public) competitors, but comparable to other private universities. There are no locations with functional monopolies, although there may be some instances where the university is the only for-profit university to offer a particular major. The university offers many services, including professional training and soft-skills courses, to stand out among competitors. Because of the steep and growing competition between for-profit universities, this paper treats the university as one of several identical for-profit university options. This paper does not consider the university-choice function of the students, but rather their choice of major within the university.

1.4 Data

1.4.1 Preferences Evaluation

The questions used in this study evaluate the motivation of individuals in their work. They were designed by the author with support of two US-based, Spanish speaking psychometricians; a Peruvian Psychologist; two education-in-development specialists; and a Mexican test specialist, as well as the staff at Servicios Farolito.[5] The team also evaluated 500-1000 responses in half of the questions to ensure that responses to each question varied sufficiently to provide useful results (internal validity check). The team developed the

[5]Kevin Joldersma, Edmond Gaible, Mary Burns, and Adriana González. For Bios, see the appendix.

questions specifically for a Peruvian context. Specifically, it is designed for Early Career positions, meaning people who are entering their first job or within 7 years of their first job. This includes salaries of between minimum wage (S/.750 per month, or $280) and a general supervisory position paying S/.2000 per month ($740).

The questions are influenced by the Effort Reward Imbalance Questionnaire (ERI). The ERI Questionnaire is used to evaluate stress at work. The ERI has been demonstrated in over 35 studies in a range of countries to predict professional stress, which often leads to health issues (Siegrist, Li and Montano, 2014). It has successfully predicted stress and poor health outcomes in Colombia, Mexico and Brazil for low-skilled workers (Ortiz, 2010).

The ERI assesses three factors: Overcommitment, Effort, and Reward. Rewards can come in three forms: Esteem, Job Security, and Job Promotion. Questions in the ERI are meant to determine how stressed the individual feels with respect to the topics covered in the test. It has been shown that the ERI is correlated with turnover intentions (Panatik et al., 2012). Given that "stress" is one of the most common reasons given for leaving a job of the type covered in this test (anecdotal), the test designers hypothesize that the ERI will be a strong predictor of turnover in the Peruvian setting.

The questions used here ask what types of rewards are more important to the respondent. Aligning the concept of reward preference with values, the proposal that these questions can identify retention or study completion is based on the ASA model, proposing that people with the same types of reward preferences will excel in similar professional settings. The set of questions used in this study translate "job esteem" to *Quality*, "job promotion" to *Achievement*, and expand the concept of "job security" to include the fact that job security comprises an important part of being able to care for one's family and spend time with one's family, *Family*. We have added *Income* as a separate factor.

In the writing of these questions, research on what makes different populations in Peru happy was also considered. The Wellbeing in Developing Countries ESRC Research Group (WeD) conducted surveys in seven villages ranging in location and level of urbanity between 2002 and 2007. The results emphasize how important social interaction is to this

particular population, with 50% of households listing it as a top motivation for decision-making (Guillen-Royo, 2008). The research also found that rural inhabitants, who had less knowledge of their relative poverty, were happier, in keeping with the Easterlin Paradox (Copestake et al., 2009). This finding informs our hypotheses about motivation by income.

Table 1.3 includes two sample questions and answers and the characteristic that they represent.

Table 1.3: *Example Test Questions*

Question / Answer	**Type**
A good professional career is one in which	
the worker has a lot of experience in many different companies.	*Achievement*
the worker stays with the same company throughout his or her life.	*Family*
the worker achieves a good salary in a short period of time.	*Income*
the worker is promoted and grows professionally within the same company.	*Quality*
The way to help my family is by	
sharing with them the important moments in life.	*Quality*
being with them as much as possible.	*Family*
supporting them economically.	*Income*
making sure they have nothing to worry about.	*Achievement*

Translated by the author.

The questions categorize the worker by the motivations that impact his or her work. There were a total of twelve questions with a four-choice multiple-choice answer. Each answer corresponds to a particular preference type. One point corresponds with one answer in favor of that type. The sum of an individual's score for each type is 12, unless the individual left questions bank. Of 880 students in this study, 826 answered all questions, 18 left one question blank, 16 left between two and 11 questions blank, and 20 students left all questions blank. A blank score is intepreted as an unwillingness to choose an option and thus a preference set not represented in these questions.

The sources of motivation are:

1. *Work quality (Quality):* a worker is committed to his or her current position and desires to be recognized for successful implementation of the tasks assigned him or her. People

with these preferences thrive in secure jobs in established companies and minimal instability. They will work long hours to meet company goals.

2. *Financial returns (Income):* a worker seeks ever-increasing income, which may include changing positions or companies in search of a higher income. People with these preferences prefer jobs that pay well.

3. *Professional achievement (Achievement):* a worker values achievement and recognition for it. He or she seek internal or external promotions, with little commitment to his or her current employer. People with these preferences thrive in managerial positions and as external consultants.

4. *Family security (Family):* a worker values a steady income and prefers a position that allows him or her leisure time with his or her family. People with these preferences prefer positions with reliable hours and pay above all else.

1.4.2 Test Implementation

For the purposes of this study these questions were included in a skills and preferences evaluation, as part of the Farolito Test, which evaluates applicants for entry-level jobs. The entire Farolito Test, including both skills testing and preference assessment, is computer-based. Participants take the test in a computer lab, supervised by a Representative from Servicios Farolito, a company developing and implementing the test for commercial purposes. This implementation was not commercial. Test-takers have up to 2 hours to complete the test. The questions are mostly multiple choice, although some math questions are open-ended. In separate implementation of only the motivation questions that does not form part of this study, participants took a total of 15 minutes to answer all 12 questions.

Challenges that may occur during the test include slow internet or computer crashes. In the case of slow internet, test-takers receive more time; Representatives from Servicios Farolito determine by how long to extend the test period at their discretion. When a computer crashes, test-takers pass to a different computer, and resume the test where they

left off. Computers freeze in 5-10% of cases, and is more likely if the computer lab is older or the test is taken during the afternoon, when bandwidth tends to be slower.

In a few implementations, a communication breakdown between the university Administration and the professors of the courses where the students took the test caused a delay in starting the test. While students were given the full 120 minutes to complete the test in all cases, a late start could have caused individual students to rush.

1.4.3 Research Sample

Between October, 2013 and June, 2014, the Farolito Test was implemented in six locations as part of an effort by the University to assess its curriculum in general and across locations and majors.

The test was offered as the final module in a professional training course that the university offers to all students. The course is a mandatory part of the students' eighth semester of study, before they begin their final year, which includes a compulsory internship. The course, and thus the test, was part of the curriculum in all but one of the locations where the university has eighth-semester students. At the time of implementation, the university had a few new locations with students who had not yet reached their eighth semester of study.

Servicios Farolito emailed the results to each student, and submitted a summary report to the university. There was no charge associated with this implementation, although the university provided the computer labs and organized the students.

Table 1.4 shows the number of students at each location in each department. In this paper, I restrict the analysis to the three departments and four locations that have more than twenty students who took the Farolito Test: Psychology, Business, and Engineering in Lima Norte, Lima Este, Trujillo, and Piura. There were other education departments that were not established in each location, which made the sample too small to separate location and career choice decisions.

Table 1.4: *Observations by Location & Department*

Department	Sum						
	Chimbote	Lima Este	Lima Norte	Piura	Trujillo	Tarapoto	Total
Business	129	106	161	151	156	106	810
Psychology	12	22	53	50	88	3	228
Engineering	168	43	127	70	124	24	557
Total	309	171	341	271	368	133	1,595

Table 1.5: *Jobs Available for Graduates by Department*

Job	Managerial?	Self-Employed / Contract Work?
Business		
Manager for industrial and commercial companies.	YES	
Manager in the finances department	YES	
Employee in business development office		
Researcher in administrative science		
Consultant in HR, marketing, productivity		YES
Consultant providing financial and risk management.		YES
Consultant training people on investment management		YES
Run own business		YES
Engineering		
Cost management for infrastructure projects		
Director or advisor to construction companies	YES	YES
Supervisor of private building projects from start to finish	YES	YES
Supervisor in Mining Industry	YES	
Supervisor of infrastructure projects	YES	YES
Psychology		
Clinical psychologist in hospitals, asylums, or rehab centers		
Private practice psychologist		YES
Research in laboratories or think tanks		
Organizational psychologist with public and private Companies		
Independent consultant		YES
Educational institutions		
Development programs with marginalized populations		

The three departments covered in this research train students for very different types of work. I reviewed two online private universities for descriptions of the types of positions that graduates from these departments can expect to secure upon graduation. These results

are described in Table 1.5. I do not have information about the frequency with which these types of jobs are actually secured by graduates in each department, but it can be expected that students who choose to pursue these studies believe that they can secure jobs in their field of study.

Table 1.6 shows student characteristics of all students included in this study, including the age and gender of the students and whether they are from households where someone has a formal job. There is no significant difference in these characteristics across location and major.

Table 1.6: *Student Characteristics*

Department	**Location**					
	Chimbote	Lima Este	Lima Norte	Piura	Trujillo	Tarapoto
Age						
Business	22.12	23.66	23.37	22.1	23.63	22.3
	(3.05)	(4.22)	(3.48)	(4.46)	(3.29)	(2.23)
Psychology	22.67	25.45	24.87	23.14	22.43	21.33
	(2.42)	(4.89)	(4.88)	(3.59)	(3.33)	(2.31)
Engineering	22.32	23.05	23.76	21.42	24.07	23.92
	(5.64)	(3.21)	(5.28)	(1.9)	(3.57)	(3.39)
Male						
Business	.33	.27	.35	.39	.31	.41
	(.47)	(.45)	(.48)	(.49)	(.46)	(.49)
Psychology	.25	.18	.21	.18	.23	.33
	(.45)	(.39)	(.41)	(.39)	(.42)	(.58)
Engineering	.77	.53	.74	.69	.65	.75
	(.42)	(.5)	(.44)	(.47)	(.48)	(.44)
Formally Employed HH Members						
Business	.7	.64	.71	.67	.64	.64
	(.46)	(.48)	(.46)	(.47)	(.48)	(.48)
Psychology	.73	.55	.79	.74	.59	.67
	(.47)	(.51)	(.41)	(.44)	(.49)	(.58)
Engineering	.73	.77	.71	.7	.65	.71
	(.44)	(.43)	(.46)	(.46)	(.48)	(.46)

The average age is 23. There is no significant difference in age by major or location. Sixty-eight percent of the students report being in the eighth semester, and 23% of students

report being in their ninth semester. Because many students are not fulltime and thus follow an unorthodox study schedule, the semester identification is not always objective; I do not consider the difference in semesters to be notable. This paper uses the number of household members with a formal job as a proxy for familiarity with formal jobs. Overall, 67.8% of students came from households where at least one person had a formal job.

Table 1.7 summarizes the preference scores in each of the four motivation variables that we look at. On average, students are most motivated by *Quality* and least motivated by *Income*. The standard deviation for each question type is around 1.5 for *Family*, *Income*, and *Quality*. At least one student answered zero questions consistent with each preference set, and at most, students answered nine of 12 questions (75%) associated with a given preference set. The standard deviation for *Achievement* is 2.77. I do not have a hypothesis for why the variance in that preference set is so much bigger. In the analysis in Section 1.6, the notable results tend to be around 0.4, in other words, less than one-third of a standard deviation. Practically, these differences are less than a one question difference; any application of these results would require a longer questionnaire to identify differences between best-fit students for each major.

In our results, the key findings are related to the *Achievement* and *Quality* answers. However, scores for each of the four characteristics are included throughout the paper because the score in one is necessarily tied to another given the zero-sum nature of the scoring.

Table 1.7: *Test Results*

Preference	**Statistic**					
	Mean	Standard Deviation	Min.	25th perc.	75th perc.	Max.
Motivation Family	2.70	1.40	0	2	4	9
Motivaton Income	2.38	1.51	0	1	3	9
Motivation Achievement	2.77	1.45	0	2	4	8
Motivation Quality	3.79	1.53	0	3	5	9

1.5 Hypothesis

My hypotheses derive from the motivations evaluated in the test and the description of the positions secured by graduates in each department, as described in Table 1.5.

1. On average, students studying for different careers will have different preferences. I propose that these preferences will be related to the expected tasks of the careers they have chosen, rather than the tasks of pursuing an education in the careers they have chosen. Any difference preferences by academic department notable and supports the hypothesis presented here. However, given the tasks of the careers themselves, as discussed above, I expect to see the following differences:

   (a) Psychology students have a significantly higher preference for work *Quality* than students in the other majors. The jobs available to Psychology majors tend to be practitioner positions in established companies, a characteristic that is attractive to people who prioritize *Quality*.

   (b) Psychology students have a significantly higher preference for *Family* than Business or Engineering majors. Positions available to psychology majors have more predictable hours and vacations, which are highly valued by populations that prioritize *Family*.

   (c) Business and Engineering students have a significantly higher preference for *Achievement* than Psychology students. The jobs available to Business and Engineering students are more likely to be consulting jobs or managerial jobs, which are motivations for people who are driven by *Achievement*.

   (d) Business students have a significantly higher preference for *Income* than psychology students: Business has a reputation for generating high financial returns compared with other majors.

2. Students from households where nobody has a formal job will have made decisions about which studies to pursue under less information. This will reveal itself in three

ways in the data.

(a) Students from households where nobody has a formal job will have preferences that are less aligned with the hypotheses presented above that are proven to be true.

(b) Students from households where nobody has a formal job will have preferences that tend towards the mean for the entire population, suggesting that students from those households made less informed decisions than students from households with formal labor market participants.

(c) Students from university campuses in locations where larger percentage of the population has tertiary education will be more motivated by *Income* due to proximity to greater wealth.

I test the hypotheses above by regressing a measurement of motivation on area of study, student characteristics, and indicators of knowledge of the labor market. I run the following regression:

$$P = \alpha + \sum_{d=2}^{3} \beta_d \cdot \Delta_d + [\tau_1 \cdot \Gamma + \sum_{d=2}^{3} (\rho_d \cdot \Gamma \cdot \Delta_d)] + \sum (\sigma \cdot \lambda) + \epsilon \tag{1.1}$$

Where P represents to what extent the student is motivated by the preference in question, Δ_d represents a dummy if the student chose to study in department d (Psychology is the omitted variable), Γ represents the student's familiarity with the labor force, and λ represents student characteristics age, age-squared, and male. I use two variables to proxy for the student's familiarity with the labor force: whether or not a household member has a formal job, and the density of the workforce in the city where the student studies that has some tertiary education.

1.6 Results

I find that student preferences *align within* majors and *differ across* majors. Preferences within majors differ between students from households with a participant in the formal labor force from students from households without. These differences are systematic, not random, and are driven by individual familiarity with the formal labor force, not the industrialization level of where the student lives.

I am unable to draw conclusions about what these differences in preference sets suggest about a student's eventual professional success. Rather, I identify differences in preferences that *may* indicate eventual differences in success, since previous research has shown that people who are professionally successful have preferences and values consistent with their job.

1.6.1 Motivation Preferences by Major

I first consider whether and how motivation preferences differ across departments. I regress the preference constructs on department, controlling for age, age squared, gender, and university location, with standard errors clustered at the city level. Table 1.8 shows the results.

Any difference in preferences between students in different departments is notable. I have hypothesized that Psychology students (the omitted variable) will have a higher preference for *Quality* and lower preference for *Achievement* than Business and Engineering students. Indeed, I find that Business and Engineering students have a lower score in *Quality* by 0.52 questions, or one-third of a standard deviation. I find that Business students have a higher preference for *Achievement* than Psychology students by 0.44 questions, or about one-sixth of a standard deviation, and Engineering students have a higher preference for *Achievement* by 0.34 questions.

I hypothesized that Psychology students will have a greater preference for *Family* than Business or Engineering students. I do not find that. Rather, Business students have a significantly higher preference for *Family* than Psychology students, but the difference is

0.067 questions, or about 5% of a standard deviation.

I hypothesized that Business students will have a greater preference for *Income* than Psychology or Engineering students, but I do not find those results.

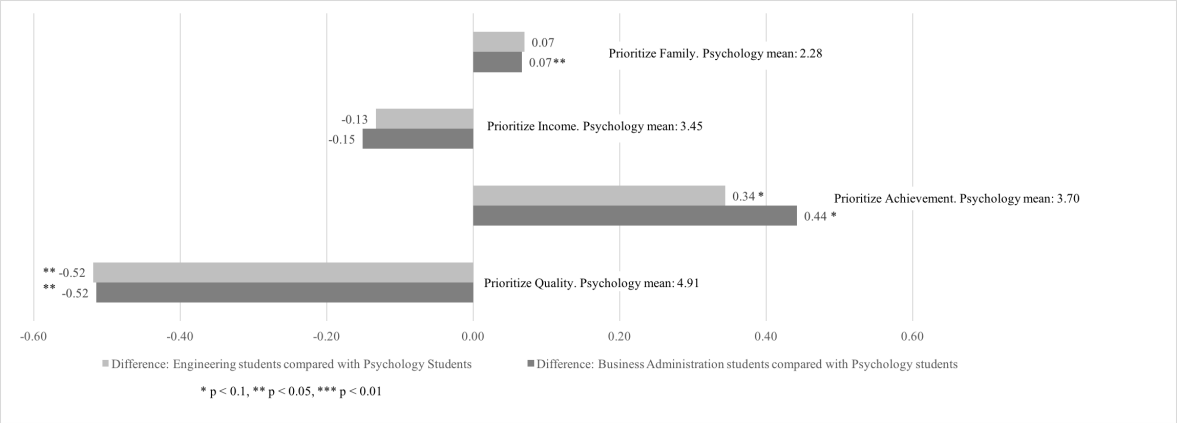
Overall, half of my hypotheses were correct. Indeed, Psychology students, who have chosen a line of study that places them in secure jobs that offer limited opportunity for management positions, do have a higher preference for secure jobs focused on practitioner work rather than management compared with students in the other two majors. Business students, who have chosen a line of study that is dedicated to preparing students for management positions, either within firms or as external consultants, do have a higher preference for recognition and promotion within an organization. These are the differences that I will examine in more detail in the next section.

Table 1.8: *Preferences and Department*

	(1)	(2)	(3)	(4)
	Family	Income	Achievement	Quality
	b/se	b/se	b/se	b/se
Business Administration	0.0666**	-0.151	0.442*	-0.515**
	(0.014)	(0.066)	(0.181)	(0.134)
Engineering	0.0699	-0.113	0.344*	-0.519***
	(0.092)	(0.169)	(0.133)	(0.043)
Constant	2.280*	3.452**	3.695**	4.907**
	(0.865)	(0.618)	(0.651)	(0.855)
R^2	0.00662	0.00622	0.0293	0.0411
N	1146	1146	1146	1146

Control for University Location, Standard Errors clustered at location level—
* p <0.10, **p <0.05, *** p <0.01

Figure 1.1: *Difference in Preferences by Department*



It is worth noting that this snapshot is eight semesters into a major. It is unclear to what extent these preferences existed when students selected into majors or were developed through the socialization of the studies. Kinsler and Pavin (Kinsler and Pavan, 2012) develop a model for the accumulation of skills over a college career, and Shwarzwaller (Schwarzweller, 1960) documents the relationship between proximate culture and the development of preferences. I take no position on this, but rather, in the next section, document the differential development of preferences between different groups.

1.6.2 Familiarity with the Formal Labor Market

I next look at whether preferences differ within majors when familiarity with the formal labor force differs.

In this study, I do not comment on whether preference convergence is due to ex-ante preferences at the start of one's university career or whether it is due to socialization throughout a career.

Household member with a formal job

In Table 1.9, I look at whether students with household members without a formal job have different preferences than their classmates within their major. Any difference in preferences

between students in the same department with a different familiarity with the labor market suggests optimization differences in career choice between students with more and less familiarity with the labor market.

Having a household member who works is an approximation of familiarity with the formal labor market. Students who grew up in households with parents or other relatives in the formal workforce who have since retired or left and students who no longer live at home will be grouped with students from households where nobody has a formal job. However, the chance that parents who once had formal jobs will have since left is small: generation intervals are relatively small, so retirement is unlikely. Given the fast growth of the formal labor sector, it is much more likely that people are entering, not leaving.

Table 1.9: *Preferences, Department, and Family Members with Formal Jobs*

	(1) Family b/se	(2) Income b/se	(3) Achievement b/se	(4) Quality b/se
Business Administration	0.109	-0.0592	0.530**	-0.820**
	(0.164)	(0.060)	(0.145)	(0.148)
Engineering	0.0697	0.0829	0.104	-0.698**
	(0.116)	(0.286)	(0.097)	(0.154)
Family Member Works	-0.0234	0.367	0.0897	-0.451**
	(0.210)	(0.181)	(0.134)	(0.124)
Business*Family Member Works^{+}	-0.0629	-0.123	-0.161	0.463***
	(0.243)	(0.173)	(0.176)	(0.036)
Engineering*Family Member Works^{+}	0.0259	-0.286	0.298	0.334
	(0.163)	(0.251)	(0.171)	(0.191)
Constant	2.198	3.337***	3.506**	5.449***
	(0.976)	(0.562)	(0.774)	(0.755)
R^2	0.00630	0.0116	0.0343	0.0416
N	1126	1126	1126	1126

Control for University Location, Age, Age^2, Gender; Standard Errors clustered at location level
* p <0.10, **p <0.05, *** p <0.01
$^{+}$ Interaction variables

I regress the preference indicators on major, a dummy variable that indicates if the student has a formally employed household member, and interaction variables of major*family member in the formal workforce. Again, I control for age, age-squared, gender, and university location, and I cluster standard errors by location.

My hypothesis predicts that the Psychology students from households with a formal labor market participant will have higher preferences for *Quality* than Engineering and Business students from households with a formal labor participant. My hypothesis predicts that these findings above will especially hold for students from households with a formally-employed member.

My hypothesis only partially holds. There are, in fact, differences in the preference sets between students from households with a formally-employed member and students from households without a formally-employed member, even within the same major. This supports the hypothesis that career optimization is different for students from these two types of households.

With respect to my specific hypotheses about the ways in which these preferences will differ, Business and Engineering students from all types of households have a lower *Quality* score than Psychology students. More surprisingly, this difference is *more pronounced* for students from households without a formally-employed worker. Business students have a *Quality* score of 0.82 less than Psychology students, and Engineering students have a score of 0.7 less than Psychology students in households without a formal labor market participant. By contrast, Psychology and Engineering students from households with a formally-employed worker are significantly less motivated by *Quality* than Psychology and Engineering students from households without a formally-employed worker (0.45), and the difference between the two groups is the same as it is for students from households without formal employees. The difference in preference for *Quality* between Business and Psychology students is 0.46 less in households with a formal employee than households without (the coefficient on *Business*Family Member Works*), and that difference is significant.

I am unable to say why my hypothesis does not result. People with preferences for *Quality* are expected to have a preference for employment within a company, but I've found that students from households without a formal employee are more motivated by *Quality*. Students from households with a formal labor market participant are more likely to pursue university education by default. Those from households without a formal labor

market participant must have a strong preference for formal employment, represented as a preference for *Quality*, in order to choose to pursue a university education. Another possibility is that, for previous generations where there were less formal employment opportunities, the people who sought formal employment were less able to self-motivate to produce high-quality work. In this way, the choices of people with a preference for *Quality* have changed as the labor market has formalized.

I turn to findings with the *Achievement* preference. My hypothesis had proposed that Psychology students will have a lower preference for *Achievement* than Business and Engineering students. In the previous analysis, that hypothesis held. In this analysis, it holds only for Business students, whose preference for *Achievement* is 0.53 higher than Psychology students. There is no significant difference between students with formally-employed household members and those without. However, the magnitude of the difference between Engineering and Psychology students without household members with formal jobs is only 0.1 (not significant) compared with a significant 0.33 in the previous section, and the difference increases by 0.3 (not significant) for households with formal labor market participants. By adding variables, I have reduced the power of the analysis, and it's possible that my hypothesis will result with a larger sample size.

I repeat this analysis replacing the dummy for having a HH member participating in the formal labor force with the number of household members with a formal job, up to two household members. This number may better capture the extent of familiarity with the labor force than the dummy variable. Dropping students with more than two household members with a formal job protects against proxying for household size with this variable. The results are in Table A.1. The magnitude of the difference in preference for *Quality* between Business/Engineering students and Psychology students increases, although it is not significant for Engineering students. Household members with a formal job are correlated with a lower *Quality* motivation score similar to that in Table 1.9. The magnitude of the difference in score for motivation by *Achievement* increases between Business students and Pschology students.

1.6.3 More about the Difference in Preferences by Family participation in the formal labor market

I hypothesized that students from households without family members in the formal labor market will be less likely to have preference sets that meet the preference-set hypotheses. I found that this is not the case.

My second hypothesis about students from households without family members in the formal labor market is that their preference sets will tend towards the mean compared with students from households with family members in the formal labor market. In other words, the differences between students in different departments would be smaller, even if they still existed.

This is also not the case. I review the differences in preference sets between students by department in separate regressions by whether or not the student has a family member with a formal job. Results are in Table 1.10. Psychology students from households without a formal labor market participant are more motivated by *Quality* than psychology students from households with a formal labor market participant, and the difference between Psychology students and Business or Engineering students is more than twice as much (0.87, 0.77) among students from households without a formal labor market participant.

I hypothesized that Psychology students are significantly less motivated by *Achievement* than Business or Engineering students. When I split up students by whether they have a formally-employed household member, only Business students in households with a formally-employed worker are significantly different from Psychology students. The direction and size of the difference is similar to the results in Table 1.8 among students with formally-employed household members. There is no practical or significant difference between Engineering and Psychology students in households without a formally-employed household member.

The manner in which students choose their area of study is consistent with their preferences for *Achievement* and *Quality*, but this difference is much stronger among students without a household member who works formally. This is contrary to my hypothesis, given

that these students have less familiarity with the labor market.

If some students from households without formal labor market participant were making informed decisions about their majors, we would see that the distribution of preferences for work were closer to the overall mean among students from households without participants in the formal labor force. In fact, we see the opposite. Business students without household members who work formally are significantly less motivated by *Quality* than Psychology majors, and the magnitude, -0.873, is greater than the magnitude on the coefficient among students from households with participants in the formal labor market, -0.328 (and that difference is not significant). Engineering students are also significantly less likely to be motivated by *Quality* than Psychology students, and the magnitude on the coefficient for students from households without formal labor market participants (-0.767) is greater than that for students from households with formal labor market participants (-0.328). The difference is significant for both groups. We see a similar pattern for *Achievement*, among business majors, students from a household without formal labor force participants have 0.530 questions greater interest (significant) than psychology majors, and that difference is 0.353 and not significant for Business students from households with formal labor force participants. There was nothing notable among motivation by *Family* or *Income*.

Table 1.10: *Preferences and Department, by Employment Status of Household Members*

	Has HH Member with Formal Job				No HH Member has Formal Job			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Family	Income	Achievement	Quality	Family	Income	Achievement	Quality
	b/se	b/se	b/se	b/se	b/se	b/se	b/se	b/se
Business Administration	0.0567	-0.197	0.353	-0.328	0.140	-0.0337	0.530**	-0.873***
	(0.091)	(0.121)	(0.244)	(0.141)	(0.161)	(0.115)	(0.130)	(0.137)
Engineering	0.0864	-0.214	0.381	-0.328**	0.172	0.0698	0.0934	-0.767**
	(0.124)	(0.182)	(0.208)	(0.077)	(0.131)	(0.362)	(0.161)	(0.218)
Constant	1.875*	4.121***	3.862***	4.240**	2.806	2.124**	2.783*	6.964***
	(0.669)	(0.580)	(0.586)	(1.155)	(1.837)	(0.567)	(0.934)	(0.590)
R^2	0.0118	0.0111	0.0344	0.0213	0.0169	0.0193	0.0515	0.103
N	761	761	761	761	365	365	365	365

Control for University Location, Age, Age^2, Gender; Standard Errors clustered at location level—
* p <0.10, **p <0.05, *** p <0.01

Because their preferences follow a normal distribution with differing means by major, and do not show indications of mean reversion (compared with the overall averages), students from households without participants in the formal labor force are sorting systematically across majors, even though they demonstrate different preferences from their classmates.

1.6.4 Are these Differences in Preferences Related to Overall Cultural Norms or are they Family-Specific?

I next consider that these findings are colinear with the general industrialization of the region where the students live, and reflect the difference in norms inherent in areas that are more or less industrialized where having a family member who works is related to the general formalization of the region. I replace the dummy of whether a family member has a formal job with a variable representing the percentage of the formal labor force with some tertiary education in the city where the student studies.

The density of workforce participation with higher education is the greatest in Lima, which applies to both Lima locations of the university. Higher-education penetration in 2008 was 38.4% in Lima, 17.7% in Trujillo, and 15.6% in Piura (Ministerio de Trabajo del Peru Dataset, 2012).

I hypothesize that higher-education penetration will not be indicative of preference-sets in the same way that having a family member who works is. I also hypothesize that an increasing percentage of higher-educated employees, as a proxy for the professional industrialization of the population, will be correlated with increased motivation by *Income* among all populations.

Table 1.11 shows the results of regressing preferences on department, on higher-education penetration, and on the interaction between the two. The variable for higher-education penetration drops out, which may be due to collinearity between the interaction variable and the location dummies.

I then include both whether a household member has a formal job and the percentage for formal workers with tertiary education and their interactions with majors. The findings

from the earlier regressions with household members participating in the labor force hold: Psychology and Engineering students are significantly more motivated by *Quality* than Psychology students. Students from houseolds with a formal employee are less motivated by *Quality* and the difference between Psychology students and Business/Engineering students with respect to *Quality* preferences decrease.

All differences in motivation by *Achievement* become insignificant, but increasing higher-education density is correlated to a greater increase in motivation by *Achievement* for Engineering and Business students than for Psychology students.

As predicted, an increase in the higher-education density is significantly correlated with an increase in motivation by *Income*. Someone in Piura, with a higher-education density of 15.6%, would have an *Achievement* score of 0.345 less than someone in Lima, with a higher-education density of 38.4%.

Table 1.11: *Preferences, Department, and Local Density of Higher-Educated Workers*

	(1) Family b/se	(2) Income b/se	(3) Achievement b/se	(4) Quality b/se
Business Administration	0.060	0.287***	-0.096	0.218
	(0.090)	(0.041)	(0.781)	(0.434)
Engineering	-0.316	-0.480	0.099	-0.368
	(0.163)	(0.668)	(0.558)	(0.239)
Percent Workers with Higher Education	-0.010	.	.	.
	(0.004)	.	.	.
Business*Perc. Workers Higher Education	0.000	-0.015***	0.018	-0.025
	(0.003)	(0.002)	(0.021)	(0.014)
Engineering*Perc.Workers Higher Education	0.013*	0.012	0.009	-0.006
	(0.004)	(0.018)	(0.018)	(0.009)
Constant	2.493**	3.154**	4.476***	4.189**
	(0.741)	(0.641)	(0.671)	(0.861)
R^2	0.008	0.009	0.031	0.044
N	1146	1146	1146	1146

Control for University Location, Age, Age^2, Gender; Standard Errors clustered at location level—
* p <0.10, **p <0.05, *** p <0.01

Table 1.12: *Major, Indicators of Relationship to Formal Labor Force*

	(1) Family b/se	(2) Income b/se	(3) Achievement b/se	(4) Quality b/se
Business Administration	0.181	0.345	-0.047	-0.093
	(0.105)	(0.152)	(0.810)	(0.411)
Engineering	-0.156	-0.309	-0.142	-0.501***
	(0.125)	(0.700)	(0.599)	(0.077)
Family Member Works	-0.021	0.370	0.097	-0.458**
	(0.213)	(0.184)	(0.118)	(0.120)
Business*Family Member Works	-0.064	-0.120	-0.171	0.475***
	(0.246)	(0.172)	(0.185)	(0.054)
Engineering*Family Member Works	0.017	-0.306	0.296	0.333
	(0.169)	(0.248)	(0.167)	(0.175)
Percent Workers with Higher Education	-0.009	0.015**	-0.030	0.042**
	(0.004)	(0.003)	(0.020)	(0.011)
Business*Perc. Workers Higher Education	-0.002	-0.013**	0.020	-0.025
	(0.003)	(0.004)	(0.024)	(0.014)
Engineering*Perc.Workers Higher Education	0.008	0.013	0.009	-0.007
	(0.005)	(0.018)	(0.022)	(0.006)
Constant	2.515*	2.760**	4.482***	4.018**
	(0.837)	(0.583)	(0.708)	(0.691)
R^2	0.007	0.015	0.036	0.044
N	1126	1126	1126	1126

Control for University Location, Age, Age^2, Gender; Standard Errors clustered at location level—
* p <0.10, **p <0.05, *** p <0.01

1.7 Conclusion

In this paper, I look at professional preferences and motivations of university students in regions of Peru that are experiencing a fast expansion of the formal labor force. I demonstrate that preferences differ between students in different majors. Furthermore, students from households with a formal labor participant do, in some cases, have different preference sets than students from households without a formal labor participant.

These findings suggest that students from households without a formal labor market participant are making informed, but different, decisions about their education than students from households with a formal labor market participant. Given the challenges that students with less familiarity with the formal labor market face successfully integrating into the formal labor market, it is worth understanding better how this decision-making differs

between the two groups and what impact it may have on their long-term professional success.

Chapter 2

Gaming the System: Do Proxy Means Tests Change Household Decision-Making?

2.1 Introduction

Increasingly, governments in developing countries look to proxy means tests (PMTs) to discern between poor and very poor for purposes of targeting social programs. Research demonstrates that PMTs are more effective than typical income tests at determining poverty levels among poor populations and the (unproven) belief is that they are complex enough to prevent households from looking like they merit access to a social program when they do not. However, these findings are based on first-time uses of PMTs in the setting being studied. Little is known about how households may take action to manipulate their PMT score in a dynamic setting. In this paper, I propose that a dynamic use of PMTs may encourage households to take action to appear more poor and access social programs. Over time, the precision-from-targeting benefits of a PMT erode, and household utility is compromised. I test this empirically by looking at Mexico's conditional cash transfer program, Progresa. I find mixed evidence of systematic manipulation: households in the

treatment group, who had more information about the PMT cutoff point, were not less likely to own assets, spend their money on non-PMT items, or improve their homes. They did spend a higher portion of their income on consumable goods, suggesting distortionary spending. Eligible households in the treatment villages were also more likely to increase the number of Progresa-eligible-aged children in their home. These findings suggest that long-run implementation of PMTs may be less successful at targeting populations than the single-use settings in which they have typically been tested. This question merits future research.

Ongoing assessments of how CCT recipients used the money compared with how they would have used the money were they not trying to stay eligible for the CCT are further confounded by the increase in income of the eligible treatment households. I resolve this by evaluating how spending allocation changes as households are closer to the eligibility cutoff on either side of the cutoff.

This paper provides two contributions to existing research: First, I establish the possibility that households distort spending in response to PMTs to evaluate eligibility for social programs. Second, propose a method of looking for spending changes in a PMT application when external factors, such as a wealth effect of the transfer, confound a simple comparison: I look at proximity to the cutoff, and indicators of increased spending on non-assets.

2.1.1 A Note on Semantics

In reading this paper, it is important to differentiate between three types of wellbeing that might be used interchangeably in common parlance. *Wealth* refers most directly to wellbeing. It can be interpreted as the expected purchasing power of an individual over the course of his or her lifetime. *Income* refers to the earnings of that individual within a given time period. *PMT Score* refers to the attempt to identify the wealth level of households through analysis of other household characteristics, including asset ownership.

These three concepts are often highly correlated, but are not identical. For example, a household that has recently begun to receive a monthly transfer such as that offered by

Progresa has a higher monthly income than it did before the transfers began. The launch of Progresa boosted Progresa households into a higher income group. The households in the income group to which these households now belong have historically earned more than the Progresa-recipient households with a transfer-inclusive income at that level. Thus, control group households as compared to Progresa-recipient households with the same transfer-inclusive income will have purchased more assets, invested more in education and health, and have a higher PMT score. Alternatively, two households with identical incomes or wealth levels may choose to spend their money differently from one another, leading to PMT-score differences that reflect nothing more than preference variations.

This paper will frequently use the terms *treatment* and *eligible.* While these terms are often synonymous, they refer to different populations in this paper. *Treatment* households are located in villages that have been randomized into the treatment group. *Eligible* households are those that are deemed poor enough to receive the program. Only *eligible* households in *treatment* villages actually receive the transfers, although *ineligible* households in *treatment* villages may also be affected by pressures to misrepresent wealth according to the PMT score. *Eligible* households in *control* villages do not receive the program during the period used in this study, but they are on track to receive the program once the pilot phase ends.

The rest of the paper continues as follows: Section 2.2 reviews literature. In Section 2.3, I develop a model to describe how households make distortionary decisions in a PMT environment. Section 2.4 describes how Progresa was implemented and Section 2.5 describes identification challenges and findings. Section 2.6 concludes.

2.2 Relevant Literature

PMTs assign values to asset ownership and other household characteristics such as access to electricity to develop a score that indicates a wealth level. Proponents of PMTs argue that PMTs are more effective at identifying the poor in developing countries because the majority of these populations work informally and inconsistently enough that documenting income over stretches of time is nearly impossible. PMTs are considered more effective because

assets and other household characteristics better indicate wealth than reported income in informal and inconsistent work environments (Skoufias et al., 2001).

Key papers about the principles of targeting in general include Besley and Kanbur (Besley and Kanbur, 1993), which outlines the considerations and goals of targeting, and Nichols and Zeckhauser (Nichols and Zeckhauser, 1982) who discuss the benefits of targeting a social program not only towards those in need, but those who are more likely to benefit from the program.

Findings in support of PMTs, however, may be inaccurate. Research supporting the efficacy of PMTs is based on the first-time use of PMTs, thus ignoring the possibility that the ineligible or borderline population could learn the PMT algorithm over time and eventually take action to reduce their PMT score. There is evidence that as the PMT score becomes known, corruption occurs at the administrative level: Camacho and Canover find that, in Colombia, politicians misreport household PMT scores as a way of purchasing patronage from their constituency (Camacho and Conover, 2011).

Alatas et al. provide evidence to suggest that although PMT usage is more effective at identifying wealth than means testing, there are alternative methods of identifying the poor which may be more acceptable to the population in question. In a randomized control trial in Indonesia, the authors establish that PMTs identify income-poor households better than community targeting, but community targeting identifies households that better meet the community's definition of poor. As a result communities are more satisfied with the fairness of their own choices than with PMT usage (Alatas et al., 2010).

To the extent that a PMT may distort spending and investment, it is similar to a tax. There is a large body of research about optimal taxation and distortionary spending around inflection points. Saez finds that US taxpayers bunch at the first kink in the tax schedule (Saez, 2010) in response to an aversion to increased taxes. Chetty (Chetty, 2012) discusses how to evaluate price elasticities when distortionary spending incentives are in place. It has been established that tax evasion is more common for hideable expenditures or sources of income (Slemrod, 2007). Lee and Card (Lee and Card, 2008) introduce the use of regression

discontinuity, similar to the analysis I use here, under conditions of specification error.

The literature about Conditional Cost Transfers as social program abound. Programs in Nicaragua (Maluccio et al., 2005), Brazil (Bourguignon, Ferreira and Leite, 2002) and Jamaica (Levy and Ohls, 2010) have found generally positive impact of differing degrees. Papers on Progresa have found positive impacts on school attendence (De Janvry et al., 2006), (Schultz, 2004), (Skoufias et al., 2001), and household bargaining and investment in family products (Attanasio and Lechene, 2010).

2.3 A Model of Distortionary Behavior

The utility optimization that leads to distortions is fairly straightforward. I assume a unitary decision-making model that was unchanged by the launch of Progresa. I assume that the net benefit of receiving Progresa in exchange for meeting the health and education requirements to receive Progresa is positive. Let a represent new assets or other indicators of wealth included in PMT score (non-consumable goods), x represent other (primarily consumable) goods, y represent household income exclusive of the Progresa transfer, T represent Progresa transfer, and β_i represent a household-specific parameter of preferences over goods.

I assign no price values to the items above, but rather normalize investments in each of these assets to 1, which could represent price. A blender may be 20 units of a while a water heater is 150 units, for example.

I begin by assuming perfect information about the PMT score and eligibility cutoff. In a single-period game or a world without functioning credit markets, an individual maximizes:

$$U_i = a^{\beta_i} x^{1-\beta_i} \tag{2.1}$$

Where ownership of all goods in period i is a function of involvement and ownership in the previous period:

$$y_i + a_{i,t-1} = a_i + x_i \tag{2.2}$$

The left-hand side of the budget constraint comprises both income, y, and the opportunity

cost of not selling assets, $a_{i,t-1}$. In a world without PMT scores, optimizing households will choose $x^* = (1-\beta_i)(y + a_{i,t-1})$ and $a^* = \beta_i(y + a_{i,t-1})$.

Now, introduce Progresa and its associated PMT eligibility identification mechanism. Households will now receive a net transfer, T, if they maintain a PMT score below the cutoff, $a_i \leq \bar{a}$ and meet the conditionality requirements. This analysis focuses only on households for whom T is positive. Households will distort asset ownership to access the program if

$$\bar{a}^{\beta_i}(y + T + a_{i,t-1} - a)^{1-\beta_i} \geq \beta_i^{\beta_i}(1-\beta_i)^{1-\beta_i}(y + a_{i,t-1}) \tag{2.3}$$

and

$$\beta_i(y + a_{i,t-1}) \geq \bar{a} \tag{2.4}$$

Alternatively, households may choose to hide assets for a cost $h(a^h)$ such that $h(a)$ is nonlinear and nondecreasing on a and eventually reaches an infinite value in that it is potentially impossible to hide a water heater. Households will now choose a combination of distorting and hiding purchases when:

$$\max_{a^h_{i,t}, a^{nh}_{i,t}} \left\{ \left((a^h_{i,t} + a^{nh}_{i,t})^{\beta_i}(y + a_{i,t-1} + T - h(a^h_{i,t})^{1-\beta_i})^{1-\beta_i} \right) \geq \beta_i^{\beta_i}(1-\beta_i)^{1-\beta_i}(y + a_{i,t-1}) \right\} \tag{2.5}$$

Hiding and distorting are corner solutions. Under perfect information, households will not choose to distort or hide more than they need to.

Progresa eligibility was imperfectly implemented, causing these optimizations to be probabilistic, which decreases the incentives to distort. A model incorporating those outcomes is included in the appendix.

The decision making summarized above highlights the following outcomes:

- Households with higher β_i preferences for assets will be less likely to distort, as the utility cost to not owning their preferred asset bundle increases.
- As the utility loss from reaching the asset ceiling increases, households will be less likely to distort. The utility loss increases as the non-PMT asset bundle increases above

$\bar{a}$.

- For households with a low h, households are more likely to hide than to distort.
- In a probabilistic setting, households with a greater chance of changing their eligibility are more likely to distort or hide. In other words, as $\pi'(a^*(\beta_i, y, a_{i,t-1}))$ increases, households will be more likely to take action to change their PMT score,

2.4 Progresa Implementation

Progresa was a conditional cash transfer program implemented in Mexico from 1998-2000, at which point it was expanded and renamed Oportunidades. The program offered bimonthly cash transfers equal to about 44% of a male day-laborers income to women in eligible households, conditional upon their enrolling their children in school and meeting health requirements such as vaccinations and regular checkups (Schultz, 2004). The program has gained renown and been used in numerous studies because it was implemented as a randomized control trial, in which the control and treatment groups were randomly assigned within the pre-identified target villages. Within control and treatment villages, some households were identified as eligible, and some as ineligible, based on their PMT score.

Baseline surveys were conducted in all households in the 506 target villages in October, 1997. Table 2.1 summarizes households included in the baseline data, and Table 2.2 shows the differences between treatment and control for relevant characteristics. The sample includes 128,177 individuals in 25,873 households. Survey information collected was used to ensure that the randomly-assigned control and treatment groups were balanced, with a few exceptions that suggest that the control group was slightly wealthier than the treatment group: the control group has higher per capita income and is more likely to own a blender and a television. However, there is no significant difference in PMT score or the portion of households identified as eligible. When treatment households first learned of their eligibility status, at some point between December 1997 and January 1998, they also knew that their

eligibility would continue through November 1999. Control households did not know their status.

Table 2.1: *Data Overview at Baseline, 1997*

	Mean	Observations	Standard Deviation
Individuals	128177		
Households	25873		
HH per Village	73 .088	25873	48.121
Members per HH	5.254	24395	2.609
Perc Eligible Population per Village	0.679	25873	0.170

Note: Includes Households without recorded PMT score

Table 2.2: *Randomization Check from Baseline Data, 1997*

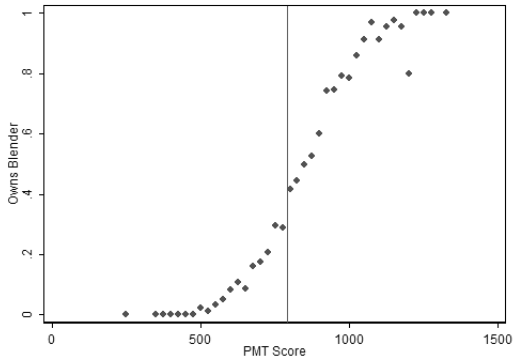
	Control Mean	Treatment Mean	Difference (Treat - Control)
Demographics			
Number of Villages	186	320	
Households per Village	73.482	72.739	-0.753
	[44.572]	[49.567]	(7.551)
Members per Household	5.393	5.318	-0.075
	[2.611]	[2.577]	(0.076)
PMT Score	795.105	737.160	-57.945
	[723.510]	[279.452]	(40.345)
Eligible	0.786	0.789	0.003
	[0.410]	[0.408]	(0.016)
Log per capita Income	3.293	3.127	-0.166**
	[1.524]	[1.650]	(0.068)
Asset Ownership			
Owns Blender	0.368	0.314	-0.055**
	[0.482]	[0.464]	(0.025)
Owns Fridge	0.165	0.142	-0.022
	[0.371]	[0.349]	(0.016)
Owns Gas Stove	0.314	0.288	-0.026
	[0.464]	[0.453]	(0.031)
Owns Fan	0.097	0.069	-0.027
	[0.295]	[0.254]	(0.017)
Owns Car	0.023	0.020	-0.003
	[0.150]	[0.140]	(0.004)
Owns Truck	0.076	0.067	-0.009
	[0.266]	[0.251]	(0.009)
Owns TV	0.500	0.447	-0.053**
	[0.500]	[0.497]	(0.027)
Owns Radio	0.656	0.631	-0.024
	[0.475]	[0.482]	(0.017)
Owns Water Heater	0.028	0.028	0.000
	[0.164]	[0.165]	(0.004)
Owns Washing Machine	0.049	0.046	-0.003
	[0.216]	[0.210]	(0.009)
Household Characteristics			
Dirt Floor	0.589	0.588	-0.001
	[0.492]	[0.492]	(0.029)
Cement Floor	0.385	0.391	0.007
	[0.487]	[0.488]	(0.028)
Straw Roof or Worse	0.727	0.716	-0.011
	[0.446]	[0.451]	(0.027)
Straw Roof	0.091	0.120	0.029
	[0.288]	[0.325]	(0.022)
Cement Roof	0.145	0.140	-0.006
	[0.353]	[0.347]	(0.019)
Cement Roof or Better	0.182	0.164	-0.018
	[0.386]	[0.370]	(0.019)
Eligible Child Lives in HH			
At Least One Eligible Child	0.7692	0.7624	-0.0068
	[0.4186]	[0.4229]	(0.0103)
Number of Eligible Children	0.7692	0.7624	-0.0068
	[0.4186]	[0.4229]	(0.0103)

Standard Deviations in brackets
SEs in parentheses from regressing outcome variable on treat by household, clustering at the village level
* p <0.10, * *p <0.05, *** p <0.01

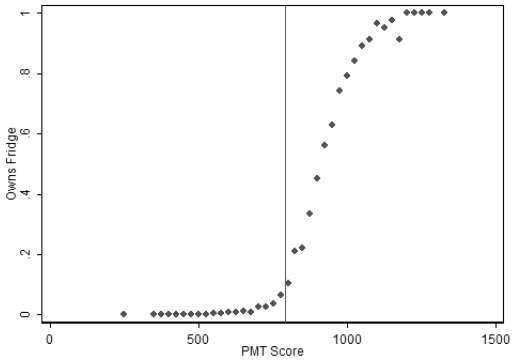
The household surveys provided the information used to both determine the PMT algo-

rithm and assign PMT scores to households. The algorithm and score differed across states, but in general it included information about the socio-economic status of the household such as portion of wage earners, children, children who work, and education of household head; household characteristics such as floor, and plumbing conditions; and ownership of assets such as televisions, radio, gas stoves, and other goods. Ownership of assets included in the PMT tend to follow an S-shaped curve with relation to PMT score, with the PMT cutoff falling somewhere in the steep part of the curve. Figure 2.4 demonstrates this relationship for one of the regions in Progresa at the baseline survey. Each dot represents a bin mean over a 10-point PMT score spread, where 1 represents ownership of the asset, and 0 represents non-ownership. As might be expected, the S-shape is stronger for more affordable assets such as blenders, and less strong for assets, such as water heaters, that even relatively wealthier populations in these generally poor, rural areas are unlikely to own.

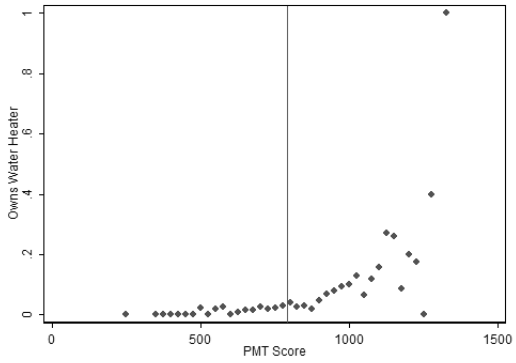
Figure 2.1: *Asset Ownership and PMT Score*



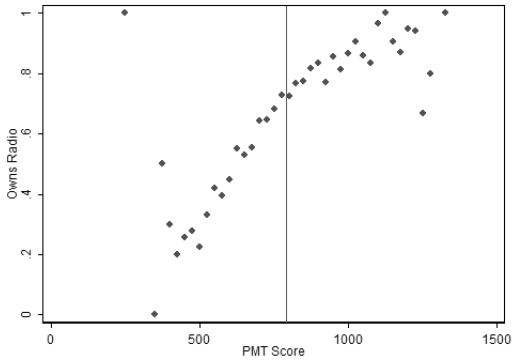
(a) Blender



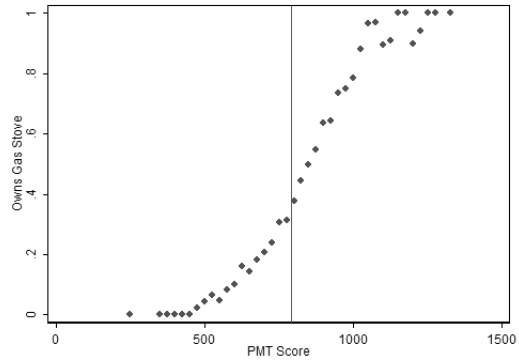
(b) Refrigerator



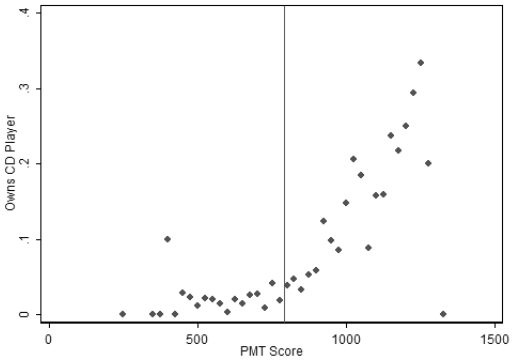
(c) Water Heater



(d) Radio



(e) Gas Stove



(f) CD Player

Note: Eligibility cutoff varied by region of implementation. Graphs above are for one of nine regions. The vertical lines denote the imputed eligiblity cutoff for that region, 791.

The original PMT score identified 52% of the surveyed households as eligible. However, the cutoff line was adjusted before the rollout of Progresa and 78.3% of households were identified as poor by the final eligibility cutoff. Figure 2.3 shows the density of households by PMT score. The eligibility cutoff is just past the peak density. Due to an administrative error, only 60% of the reclassified treatment households received Progresa transfers during the initial rollout (Angelucci and De Giorgi, 2009); Because I cannot confirm that irregular transfer allocation is actually random as opposed to systematic, I consider all eligible households to be eligible, although some may have been misinformed as to their eligibility.

Figure 2.2: *Portion of Population Categorized as Eligible by Distance from Cutoff in PMT Score (10-PMT-point bins)*

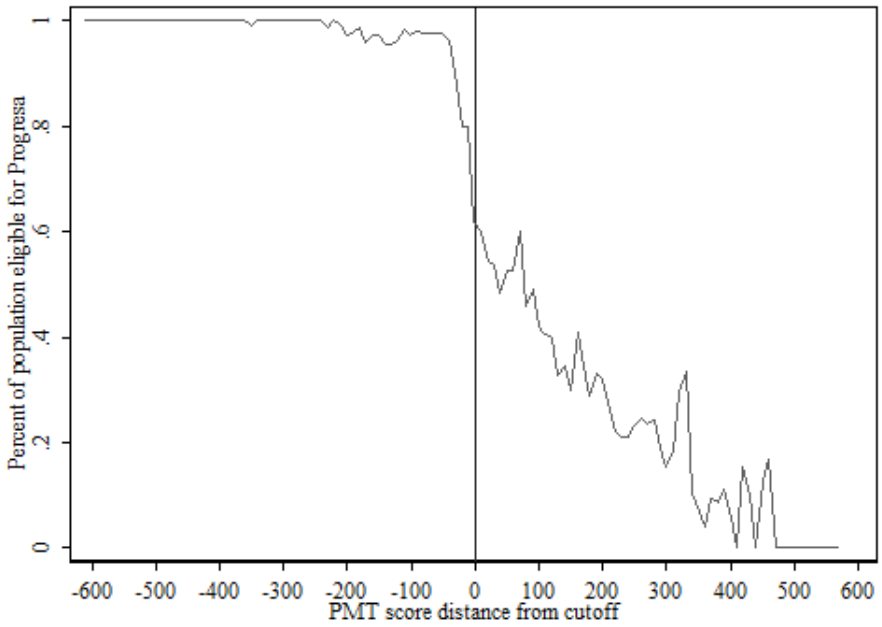
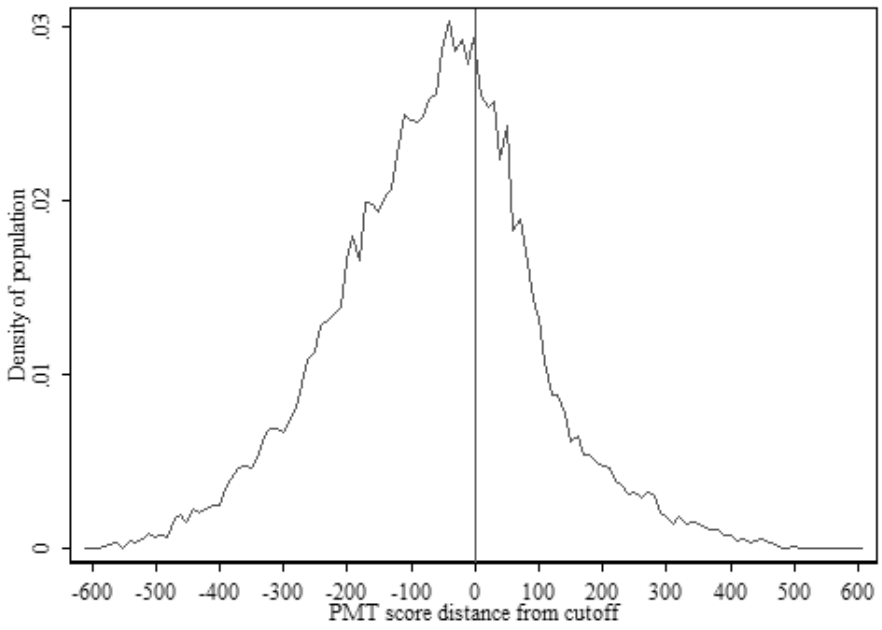


Figure 2.3: *Population Density by Distance from Cutoff in PMT Score*



A second baseline survey was implemented in March, 1998, and Progresa made its first transfers that month. Bimonthly transfers continued until the program expanded to the control group in November, 1999. During the period of implementation, household surveys were implemented in the control and treatment groups in October 1998, March 1999, and November 1999. Baseline and followup are connected with a household individual identifier. I verify that households are correctly matched and information is consistent over time by looking at the change in reported age of the same household members between the baseline and follow-up. Only two percent of individuals report having an age that is different from one or two years older than the age they reported being eighteen months ago. There is no evidence of teenagers about to become too old to be eligible for Progresa transfers being more likely to incorrectly report their age than other children. I conclude that household characteristics are reliably consistent between baseline and follow-up.

In my analysis, I use November 1997 data as the baseline data on all topics. Outcome variables are constructed as described in the analysis section.

2.5 Empirical Analysis

2.5.1 Identification and Associated Challenges

I now turn to an application of this model using Mexico's Progresa program. To test the hypotheses presented above, I turn to the implementation of Progresa for several reasons. First, Progresa was a randomly-implemented program in which both control and treatment groups received a PMT score, but only members of the treatment group learned their eligibility outcome. The control group is a useful control for the effects of *knowing* eligibility. Second, the concept of a PMT score was unique enough that households in the study did not anticipate PMT usage and thus manipulate their PMT score ahead of the baseline program. Finally, unlike other well-documented, PMT-employing programs that targeted only female children or orphans, the program was available to all eligible households in the treatment villages with children. This generates a large dataset of 24,000 households, which is diverse and thus more conducive to external validity. However, the Progresa dataset is not ideal for several reasons. Ultimately, the confounders make identification impossible. In this section, I discuss the confounding aspects of the Progresa dataset and describe how my identification strategy aims to get around it.

Wealth Effect

This analysis relies on comparisons between the control and treatment group to predict what asset ownership the treatment group would have in the absence of the incentive effect generated by PMT use. However, the treatment group also benefits from the wealth effect of the Progresa transfers, which went only to the eligible households in the treatment villages, and have been shown to have a spillover effect on ineligible households as well (Angelucci and De Giorgi, 2009).

The goal of this analysis is to isolate the effect of PMTs on distorting or hiding. Ideally, the treatment and control groups would be identical in all respects except for the incentives to distort. However, the distortionary effect is not isolated because treatment group households

have higher (Progresa-transfer inclusive) income than control group households in the followup. They are likely to have spent some of this increased income on assets. This is the wealth effect of the program.

One way to reduce this bias is to review whether indications of distortionary spending increase along the eligibility cutoff. Distortionary spending would be higher there because the impact on eligibility would be greater. In a perfect eligibility implementation, households just above the cutoff would not have benefited from the Progresa transfers. [1] I do this in the next section.

Eligibility Manipulation

In this implementation of Progresa, the eligibility cutoff was not cleanly implemented in two ways. First, the construction of the PMT score was not well documented (email exchange with Susan Parker, March 2011). In other words the weighting of different household characteristics in the PMT algorithm is not known. It has since been redesigned. In my identification of characteristics included in the PMT score, I use a document listing a range of potential PMT algorithms by region, which appears to consider several PMT options, published internally by the *Dirección General de Planificación y Evaluación* in 2004.

Second, the eligibility cutoff was not strictly implemented in that there is no PMT score in any region below which the eligibility assignment is 100% for all households with children and above which the eligibility assignment is 0% for all households. The unclear eligibility cutoff is particularly complicated because it is known that policymakers adjusted the eligibility cutoff to include more households between November, 1997, and the launch of the program in March, 1998 (Angelucci and De Giorgi, 2009). Of those reassigned to eligible status, only 60% of the households ultimately received the transfers. The lack of a cutoff to determine eligibility combined with little information about the PMT algorithm makes it difficult to identify an eligibility cutoff, In Table 2.3, I show the relationship between

[1] There is some indication that ineligible households also benefited from Progresa transfers (Angelucci and De Giorgi, 2009) but these benefits are minimal.

identification as eligible and receipt of a Progresa transfer within the treatment group. I find that no households identified as ineligible received a transfer, but only about 73% of all households eligible for the transfers received any transfers in the pilot time period.

Table 2.3: *Eligibility Assignment and Receipt of Transfers in Treatment Group*

	Received Transfers			
Eligibility Status	Yes	No	Blank*	Total
Eligible	7767	1983	892	10642
Ineligible	0	863	2098	2961
Blank*	995	334	1044	2373
Total	8762	3180	4034	15976

*Blank transfer information likely reflects households that have never been eligible

I impute the cutoff by running regression (2.6) for every potential cutoff between the PMT score of the lowest-scoring ineligible household, and the PMT score of the highest-scoring eligible household, where I_i indicates a dummy that equals 1 if the household is ineligible, and 0 if the household is eligible; S_i represents the household's PMT score; and C_r represents the region-specific eligibility cutoff.

For each region, I use the cutoff point for which the regression yielded the highest R-squared.[2]

$$I_i = \alpha_2 + \gamma_1 \cdot S_i + \gamma_2 \cdot (S_i \geq C) + \epsilon_{2i} \tag{2.6}$$

Figure 2.2 shows the eligibility density by proximity to eligibility cutoff. I find that the imputed cutoff is quite close to the point at which eligibility is practically 100%. Yet eligibility declines slowly over the 300 points in the PMT score. Eligibility density of zero occurs only at PMT scores well above the eligibility cutoff.

The fact that households with the same PMT score have different eligibility assignments

[2]I also did this for ineligible = cutoff, ignoring PMT score. The predicted cutoff was never more than 100 points from the one used here, and in six of nine cases, it was the same.

indicates the presence of some level of either corruption, incompetence, or both. In the corrupt state, households would have been able to make transfers to decision makers in exchange for access to the program. In the incompetent state, eligibility assignment would have been haphazard for no reason other than poor oversight or lack of clarity of rules. Perhaps the eligibility cutoff was not clearly defined for the individuals determining which households were eligible, or when the eligibility cutoff was adjusted, only a portion of the households were reassigned in the database. Attempts to resolve this question through conversations with policymakers failed due to time elapsed since this implementation.

Table 2.4 examines whether some household characteristics increased a household's probability of an eligible designation, and whether being in the treatment group had differential effects on eligibility assignment among households in the area above the imputed cutoff, where the cutoff by PMT score was less strictly implemented. Column (1) shows the coefficient κ_2 from equation (2.7), where X_i represents a vector of household characteristics included in the PMT score. Column (2) shows the coefficient τ_4 in equation (2.8) to test whether eligibility was assigned differently in treatment than in control groups.

$$E = \kappa_1 + \kappa_2 \cdot \mathbf{X_i} + \epsilon \tag{2.7}$$

$$E = \tau_1 + \tau_2 \cdot \mathbf{X_i} + \tau_3 \cdot T + \tau_3 \cdot T \cdot \mathbf{X_i} + \epsilon \tag{2.8}$$

Were the eligibility cutoff perfectly implemented, we should see that, controlling for PMT score, household characteristics have no effect on eligibility. That is not the case here. I find that, controlling for proximity to the eligibility cutoff, poorer households are more likely to be identified as eligible – ownership of assets and monthly per capita income is negatively associated with eligibility, while having unschooled children or disabled individuals as households members is positively associated with eligibility.

Table 2.4: *Relationship between Characteristics, Treatment Status and Eligibility*

	(1) Eligible/Ineligible Differences	(2) Differences by Treatment Status
Treatment	-0.022*	0.027
	(0.011)	(0.084)
Distance from imputed eligibility cutoff	0.000**	-0.000**
	(0.000)	(0.000)
HH has child of eligibile age	0.063***	-0.000
	(0.014)	(0.029)
Blender	-0.025*	0.015
	(0.014)	(0.028)
Fridge	0.010	-0.012
	(0.014)	(0.028)
Gas Stove	-0.012	-0.024
	(0.014)	(0.027)
Water Heater	0.023	-0.032
	(0.025)	(0.050)
Radio	-0.058***	-0.023
	(0.014)	(0.028)
TV	-0.094***	-0.001
	(0.014)	(0.029)
Washing Machine	-0.031	-0.036
	(0.019)	(0.038)
Fan	-0.118***	0.058*
	(0.017)	(0.033)
Car	-0.368***	-0.006
	(0.013)	(0.025)
HH has disabled member	0.117***	0.032
	(0.024)	(0.048)
Dirt Floor	-0.009	0.006
	(0.031)	(0.064)
Cement Floor	-0.056*	-0.022
	(0.031)	(0.063)
Number members in household	-0.050***	0.007
	(0.004)	(0.008)
Num. members under 15 who worked	0.116***	-0.035
	(0.024)	(0.049)
Num. members under 15 not in school	0.012	0.013
	(0.020)	(0.041)
Age HH head	-0.000	-0.001
	(0.000)	(0.001)
Num. people per room in HH	0.005***	-0.001
	(0.001)	(0.001)
Ln per capita monthly hh income	-0.018***	0.002
	(0.003)	(0.006)
r2	0.145	0.147
N	7287	7330

(1) Regresses characteristics on "Eligible".
(2) Regresses characteristics and treatment interactions on "Eligible." Interactions are shown here.
Note: Robust standard errors, region FE
* p <0.10, * *p <0.05, *** p <0.01

If these differences were due to corruption, we would see evidence of households "purchasing" their way to eligibility when they were aware of their eligibility assignment. These trends are no different in treatment villages than control villages. It appears that ownership of a fan and negative distance from PMT score are more positively associated with eligibility status in the treatment group. Yet neither of these coefficients are practically significant. Given the large number of unobservables that may drive preference for access to Progresa, this evidence is at best circumstantial support against the possibility of corruption in eligibility assignment.

The imperfect eligibility assignment injects uncertainty into the decision-making on the part of households interested in accessing or maintaining Progresa eligibility in two ways. First, if households with identical PMT scores had different eligibility assignments, this would confuse households' ability to discern the relative weight of characteristics into the eligibility assessment and thus identify ones own proximity to the eligibility cutoff. Second, households now play a probabilistic game: by giving up a television, for example, households may improve their chance of getting a Progresa transfer by 20%, but whereas sure access to the program would encourage a household to forego television ownership, a 20% chance of access to the program may not be enough, particularly if the household is risk averse.

The imperfect eligibility assignment further confounds the analysis by generating a wealth effect within the segment of the population considered to be just above the cutoff. Because I cannot prove that imperfect implementation is random as opposed to systematic, I cannot restrict my analysis to only those households near the eligibility cutoff that were identified as ineligible.

The fact that some households with identical PMT scores have different eligibility assignments may be the result of corruption or incompetence. In either case, the imperfect correlation between eligibility and PMT score further complicates analysis by making the household decision a probabilistic game, and by increasing the difficulty of separating the wealth effect from the incentive effect by generating a wealth effect in a portion of

households above the imputed eligibility cutoff.

Poorly-Publicized Eligibility Assignment Mechanism

One year into the program, qualitative research in treatment villages noted that there was some confusion about how eligibility was determined within the village. Some households thought eligibility was randomly assigned within the village. This confusion was not pervasive, but was enough to cause the writers of the report to suggest that the eligibility mechanism be better publicized. Today, Mexican households are well aware of the PMT use, although the algorithm remains a mystery to most potentially eligible populations (conversations with employees of Innovations for Poverty Action, Peru; Techos para Mexico; and SEDESOL (Secretaria de desarrollo social)). In areas with extreme confusion about the eligibility-assignment mechanism, it is unlikely that households will take action to manipulate their score.

Political Realities of PMT Usage

The political realities of program eligibility, particularly in the context of Progresa, are two-fold. First, the political repercussions of rescinding program eligibility are high. Households may infer that once they are identified as poor, eligibility for a program designed to end when the youngest child in the household reaches late teenage years will not end before that. Second, Progresas pilot implementation period, 1997-1999, coincided with the end of the PRIs seventy-one year single-party rule. As PRI's political future became less certain, so did Progresa's. The optimization decision about asset acquisition and PMT score became even more probabilistic as a regime change may have caused a cancellation of Progresa. In reality, the program was renamed but remained largely unchanged, but that outcome was uncertain during the time of this study.

I examine whether households inferred that, once they began receiving the program, they would not be cut off and thus were free to invest in PMT indicators by comparing eligibility assignment changes between the control groups, who did not know their eligibility status,

and treatment groups, who did, in the Progresa pilot and its expansion as Compartamos in 2000. If reassessment were politically infeasible, no household in the treatment group would have been reassigned, but some households in the control group, who were unaware of their eligibility identification, would have been reassigned if their wealth and PMT score had changed over the previous three years. The data does not update eligibility assignment information between 1998 and 2000, however, I can look at receipts of transfers over this time period. I assume that households not included in the transfer dataset never received transfers, and show this outcome in Table 2.5. I find no difference between control and treatment in reassignment rates, suggesting that there was no political pressure to allow eligible households to maintain their status even if they became wealthier in the places where eligibility status was already revealed.

Table 2.5: *Changes in Access to Progresa Transfers*

1998 Status	**2001 Status**		
	Eligible	Ineligible	Total
Eligible	0.2316	0.0122	0.2438
Ineligible	0.2431	0.5131	0.7562
Total	0.4747	0.5253	1

There are several challenges to using the Progresa implementation environment to examine PMT-score manipulation. The wealth effect of the program bias treatment-control group comparisons, the imperfect correlation between PMT score and eligibility assignment infuses a wealth effect into the ineligible populations and caused uncertainty that may have discouraged households from trying to manipulate their score. There is also the possibility that some households did not fully understand the PMT mechanism, that eligible households faced a very low probability of losing their eligibility assignment, and that the uncertain future of the PRI and their social programs decreases distortionary benefits.

2.5.2 Analysis

This analysis looks at four indicators of distortionary behavior on the part of households in the treatment group: Asset ownership, having an eligible child in the household, reported expenditure on non-PMT-included items, and whether or not the household reports making an improvement on their home. Table 2.6 shows a simple difference between all relevant characteristics at the time of followup. The only significant difference is that the treatment group was more likely to have reported investing in floors or walls during the pilot implementation period.

Below, I present a simple difference and a difference-in-difference between the control and treatment group, and close with an analysis of household behavior by how close the household is to the eligibility cutoff. I find evidence of a wealth effect, evidence of distortionary spending away from assets, and some evidence of increasing the number of eligible-aged children in the household.

Assets

Expectations about distortionary behavior in asset investments by households in the treatment group are described above. Table 2.7 shows a difference-in-difference of asset ownership by treatment and control households at baseline and in the followup. I generate a composite variable of ownership of hideable assets, ownership of non-hideable assets, and total assets. The hideable assets variable includes reported ownership of a blender, fan, television and radio. The non-hideable assets variable includes reported ownership of a refrigerator, gas stove, water heater, and washing machine. The composite variable takes a value of 0 if the household reports owning none of the assets in question, and 1 if the household reports owning all of the assets in question, a decimal equivalent of the fraction of assets in the index reported owned otherwise. The outcomes do not change if, instead of weighting each asset equally, I weight the composite variable by the average price of these goods.

Table 2.6: *Simple Difference in Asset Ownership and Household Improvement, November 1999*

	Control Mean	Treatment Mean	Difference (Treat - Control)
Total Assets	0.250	0.237	-0.013
	[0.236]	[0.227]	(0.015)
Hideable Assets			
Owns Blender	0.333	0.309	-0.024
	[0.471]	[0.462]	(0.025)
Owns Radio	0.586	0.577	-0.011
	[0.49]	[0.494]	(0.020)
Owns CD Player	0.040	0.039	-0.001
	[0.196]	[0.193]	(0.005)
Owns TV	0.454	0.433	-0.021
	[0.498]	[0.460]	(0.029)
Owns Fan	0.101	0.080	-0.021
	[0.301]	[0.271]	(0.015)
Hideable Assets	0.303	0.288	-0.015
	[0.262]	[0.254]	(0.016)
Non-hideable Assets			
Owns Fridge	0.176	0.158	-0.018
	[0.381]	[0.364]	(0.017)
Owns Gas Stove	0.289	0.281	-0.008
	[0.453]	[0.450]	(0.030)
Owns Water Heater	0.024	0.026	0.001
	[0.154]	[0.158]	(0.004)
Non-Hideable Assets	0.163	0.155	-0.008
	[0.253]	[0.246]	(0.015)
Household Improvement+			
Floor	0.031	0.042	0.011**
	[0.172]	[0.200]	(0.005)
Roof	0.054	0.060	0.006
	[0.226]	[0.238]	(0.005)
Walls	0.020	0.025	0.005**
	[0.140]	[0.158]	(0.003)
Drain	0.003	0.004	0.001
	[0.052]	[0.062]	(0.001)
Pipes	0.008	0.005	-0.003
	[0.091]	[0.074]	(0.002)
Toilet	0.016	0.020	0.004
	[0.127]	[0.141]	(0.003)
Electricity	0.007	0.010	0.003
	[0.085]	[0.098]	(0.002)
Rooms	0.026	0.029	0.004
	[0.158]	[0.169]	(0.003)
Any Change	0.125	0.146	0.020*
	[0.331]	[0.353]	(0.011)
Total Changes	0.164	0.195	0.031**
	[0.508]	[0.560]	(0.015)
Eligible Child Lives in HH			
At Least One Eligible Child	02.296	02.310	0.014
	[01.966]	[01.943]	(0.061)
Number Eligible Children	02.2955	02.3095	0.014
	[01.9658]	[01.9425]	(0.0612)

Standard Deviations in brackets
Standard Errors in parentheses from regressing outcome variable on treat by household, clustering at the village level
+ Dummy variables = 1 if household reported making improvements to that part of home
* p <0.10, **p <0.05, *** p <0.01

I find that households in the treatment group increased their ownership of non-hideable

and hideable assets significantly more than households in the control group over this time period. This supports the wealth effect, in that households in the treatment group had more income over this period.

Table 2.7: *Difference in Difference: Assets*

	(1) Non-Hideable Assets	(2) Non-Hideable Assets	(3) Hideable Assets
	b/se	b/se	b/se
Post	-0.00245	-0.00289	0.0171***
	(0.002)	(0.002)	(0.004)
Treat*Post	0.0118***	0.00680**	0.0138**
	(0.003)	(0.003)	(0.005)
Constant	0.262***	0.160***	0.348***
	(0.001)	(0.001)	(0.001)
r2	0.00127	0.000334	0.00459
N	49567	46893	49567

Household Fixed Effects, Regressions clustered at village level.
* p <0.10, **p <0.05, *** p <0.01

I then control for PMT score and treatment assignment, looking at post-implementation asset ownership among the eligible and ineligible populations separately. Among the eligible populations, shown in Table 2.8, I would expect to see that ownership of assets increases with an increasing PMT score, but that this increase would be less steep for treatment households if they were trying to avoid becoming ineligible. Among the ineligible populations, shown in Table 2.9, I expect to see a a negative coefficient on Treatment, suggesting that ineligible households acquired assets at lower rates in the treatment areas than the control areas, but a positive coefficient on Treatment*PMTScore, suggesting that this distortion occured more among the ineligible populations closer to eligibility.

Indeed, I find that the PMT score at the start of the program is correlated with asset ownership. I also find that, controling for PMT score, on average, more assets were acquired by the end of the program among treatment households, and that this acquisition was evently divided between hideable and non-hideable assets. This finding is significant at the 10% level. However, the acquisition of assets by the treatment group decreased compared with the control group by 0.001 as the PMT score increased. At a PMT score of 510 for the

all asset bundle, the acquisition of assets for treatment groups would become *less* than the acquisition of assets by the control group. Of the eligible / ineligible cutoffs I imputed, the typical was between 600 - 800, so there would be negative asset acquisition among the population most at-risk of losing eligibility. This finding is significant at the 5%-10% level. I review this disparate asset acquisition in more detail in my assessment of asset acquisition by different groups, and am unable to replicate these results.

The ineligible population, shown in Table 2.9 also had increasing ownership of assets along the PMT score at the start of the program. However, at the end of the program, asset ownership was about 0.2 assets *less* in the treatment group than in the control group, significant at the 5% level. This is unaffected by the PMT score of the household: the coefficient on Treatment*PMT Score is 0 and not signficant. It is possible, then, that households at every PMT score that were wealthy enough at the start of the program to be ineligible, desired access to the program, and avoided acquiring assets.

Table 2.8: *PMT Score and Asset Ownership: Eligible Population*

	All Assets		Hideable Assets		Non-hideable Assets	
Treatment	-0.049	0.510*	-0.036	0.361*	-0.036	0.361*
	(0.074)	(0.288)	(0.052)	(0.187)	(0.052)	(0.187)
PMT Score	0.005***	0.005***	0.003***	0.003***	0.003***	0.003***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Treatment x PMT Score		-0.001*		-0.001**		-0.001**
		(0.000)		(0.000)		(0.000)
r2	0.225	0.226	0.164	0.165	0.164	0.165
N	16296	16296	16322	16322	16322	16322

Notes: Standard errors clustered at village level, controls for region
* p <0.10 ** p<0.05, *** p<0.01

Table 2.9: *PMT Score and Asset Ownership: Inligible Population*

	All Assets		Hideable Assets		Non-hideable Assets	
Treatment	-0.251**	0.325	-0.162**	0.192	-0.162**	0.192
	(0.105)	(0.516)	(0.068)	(0.343)	(0.068)	(0.343)
PMT Score	0.008***	0.009***	0.004***	0.005***	0.004***	0.005***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Treatment * PMT Score		-0.001		-0.000		-0.000
		(0.001)		(0.000)		(0.000)
r2	0.243	0.244	0.168	0.169	0.168	0.169
N	4386	4386	4398	4398	4398	4398

Notes: Standard errors clustered at village level, controls for region
* p <0.10 ** p<0.05, *** p<0.01

Having a Child of Eligible Age in Household

In the second column of Table 2.10 I look at whether a household has a child within the age group that makes a household eligble for Progresa transfers. It is possible, particularly during the slow expansion of Progresa, that households within the treatment area, particularly households without an eligible child, would invite cousins or other family members in other parts of the country to live with them in order to receive the transfer. I test this by looking at the probability that a household reports having a child between the ages of six and fifteen. There are no transfers given for toddlers and infants; given the length of time between birth and transfer eligibility, the incentive for households to conceive more children to access the program is low. I look at the probability of having a child of the given age in the household because it reflects changes made specifically to access the program, not a wealth effect of the program. Households already receiving Progresa transfers could invite children from ineligible households or households outside the region because they are now wealthier and thus can care for additional family members; the dummy variable captures a desire for the transfer rather than the wealth effect.

I find that the number of households with a child of transfer-eligibile age increased in both treatment and control groups between the baseline and followup, and that households in treatment villages are more likely to add a child of transfer-eligible age than control households.

Spending

Households that decide not to spend money on assets will either save it or spend money on other items. In a world of imperfect credit markets, at least some portion of the money saved from the non-purchase of an asset will go towards the purchase of another item. To this end, I examine the reported spending habits of households in the treatment group compared with households in the control group. A sign of distortions would be that treatment group households reported spending more on non-asset items than control groups.

Table 2.10: *Difference in Difference: Home Improvement and Has Eligible-Aged Child*

	(1)	(2)
	Made Home Improvement	Has Eligible-Aged Child
	b/se	b/se
Treatment	0.0210**	
	(0.011)	
Post		0.0137***
		(0.004)
Treat*Post		0.0106**
		(0.005)
Constant	0.125***	0.622***
	(0.008)	(0.001)
r2	0.000880	0.00353
N	22594	47766

Household Fixed Effects, Regressions clustered at village level.
"Home Improvements" includes reported investment in floor, roof, walls, drain, pipes, toilet, electricity, or rooms
* p <0.10, **p <0.05, *** p <0.01

To examine this, I look at the relationship between self-reported log per capita income and log per capita spending on clothing, on food, and total reported spending over the one-month period preceding followup surveys. The total expenditure section of the Progresa surveys includes food, clothing, transportation, school, medicine, kitchenware and sin goods like tobacco and alcohol. It does not include long-term household improvements or electronics, both of which are included in the PMT score. Household spending reflects about 13.5% of reported income, including Progresa. These questions are included in Table B.1 in the appendix.

To test the reliability of the self-reported earnings and expenditure, I examine reported log per capita spending on food and overall spending as a function of log per capita income. I find in Table 2.11 that the relationship is positive, as we might expect. I find that the relationship is stronger when I include the Progresa transfers in household income than when I look exclusively at non-Progresa household income. This suggests that households wrap Progresa transfers into household spending rather than earmark them for other uses. For ineligible or control populations, the log per capita income including Progresa transfers is equivalent to the log per capita income not including Progresa transfers. Poor households report spending about 12% of their income on food.

Table 2.11: *Relationship between Household Income and Spending*

	(1)	(2)	(3)	(4)
	Food		Total	
Income with Progresa	0.124***		0.135***	
	(0.005)		(0.005)	
Income without Progresa		0.033***		0.021***
		(0.004)		(0.005)
r2	0.053	0.005	0.051	0.002
N	21270	17524	21270	17524

Note: Robust Standard Errors
* p <0.10, * *p <0.05, *** p <0.01
Note: Expenditure and Income values are Log Per Capita

In Table 2.12, I regress log per capita spending on food, attire, and total spending as a function of income, including Progresa transfers and treatment assignment. I find that the treatment group spends significantly more on everything than the control group. In columns 2, 4, and 6, I address the proposal in Angleucci and Attanasio (Angelucci and Attanasio, 2009) that the change in spending is a result of women acquiring greater bargaining power in the household. If this were true, we would find that the impact of treatment on female-headed households differs significantly from the impact on households headed by a male, as households headed by men would be subject to shocks in household bargaining. While I do find differences in the way households spend money when they are headed by a woman overall, there is no additional treatment effect on female-headed

households than male-headed households. My findings thus do not support the fact that these differences in spending are the result of giving bargaining power to women, but rather the result of differences in the preferences of households with respect to spending.

Overall, treatment households spent a greater portion of their income on non-PMT items than control households, suggesting they are diverting expenditure from other items. This supports a hypothesis that treatment households distorted spending away from asset investment to access Progresa.

Table 2.12: *Effect of Treatment on Expenditure on Non-Asset Items, 1999 Follow-Up*

	(1)	(2)	(3)	(4)	(5)	(6)
	Food Expenditure		Attire Expenditure		Total Expenditure	
Income with Progresa	0.029***	0.029***	0.127***	0.127***	0.018**	0.017**
	(0.006)	(0.006)	(0.018)	(0.018)	(0.007)	(0.007)
Treatment	0.118***	0.119***	0.437***	0.443***	0.098***	0.107***
	(0.032)	(0.032)	(0.091)	(0.092)	(0.035)	(0.034)
Female HH Head		0.240***		-0.262***		0.347***
		(0.031)		(0.092)		(0.041)
Female*Treat		-0.001		-0.063		-0.071
		(0.040)		(0.116)		(0.051)
r2	0.013	0.027	0.021	0.024	0.006	0.025
N	17524	17520	17524	17520	17524	17520

Note: Standard errors clustered at village level
* p <0.10, * *p <0.05, *** p <0.01

Household Improvements

The PMT score includes indicators relating to the materials of the household floors, plumbing conditions, and other characteristics. The example PMT scores examined for this research suggested that the weighting of these indicators and their inclusion in household PMT scores varied widely across states. Follow-up surveys did not include information about housing-structure characteristics as the baseline survey did. However, each survey throughout the Progresa pilot included a question about whether the household made an improvement on different parts of their home since the previous survey. A list of the household characteristics covered in the survey is included in Table B.1 in the appendix. In this analysis, I generate a

dummy variable for whether the household reported making a home improvement during any survey between the baseline and final followup. The results are similar to when I conduct the same analysis looking at the number of improvements made by the household between the baseline and the final followup. I find that treatment households were more likely to make housing-structure improvements than the control group.

2.5.3 Distortionary Behavior by Distance from Cutoff

The model predicts that distortions will be greatest among the population closest to the eligibility cutoff. The different eligibility cutoffs in each region, the predicted U-shape of the distortion, and imperfect implementation of eligibility cutoffs complicate the use of raw PMT scores to identify predicted distortions. For this reason, it is useful to examine the difference in outcomes between control and treatment group by proximity to eligibility cutoff.

In Figure 2.4, I show the coefficients and standard errors of σ in Regression (2.9) and (2.10), where variables are consistent with previous equations and P is a dummy variable that takes the value 1 if the data is from the followup surveys. Regression (2.9) looks exclusively at followup data and applies to assets and eligible-aged children, and regression (2.10) is time series data that includes baseline observations and applies to spending and household improvements.

$$\begin{aligned} A_i =& \alpha_3 + \sum_{j=-7}^{j=5} \eta_{1j}(C_r + 100j) \leq S_i \leq (C_r + 100(j+1)) \\ &+ T_i \cdot \sum_{j=-7}^{j=5} \sigma_{1j}(C_r + 100j) \leq S_i \leq (C_r + 100(j+1)) + \epsilon_{3i} \end{aligned} \tag{2.9}$$

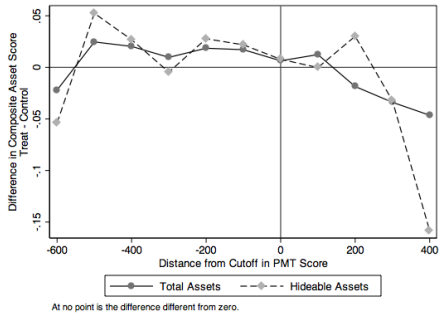
$$\begin{aligned} A_i =& \alpha_3 + \eta_2 \cdot T + B \cdot P \\ &+ T_i \cdot P \cdot \sum_{j=-7}^{j=5} \sigma_j(C_r + 100j) \leq S_i \leq (C_r + 100(j+1)) + \epsilon_{3i} \end{aligned} \tag{2.10}$$

Note that this regression does not omit any PMT-score bin, and thus drops the control for Treatment in equation (2.9), and for Treatment*Post in (2.10). This is to highlight any differences between the control and treatment group at varying proximities to the eligibility cutoff, rather than differences between members of the treatment group at different distances from eligibility cutoff. I will address the question of differential responses to the PMT score by the treatment group depending on distance from eligibility cutoff below. I restrict my analysis to households with PMT scores 600 points below the eligibility cutoff and 500 points above the cutoff because the number of households outside that range is too small for this analysis.

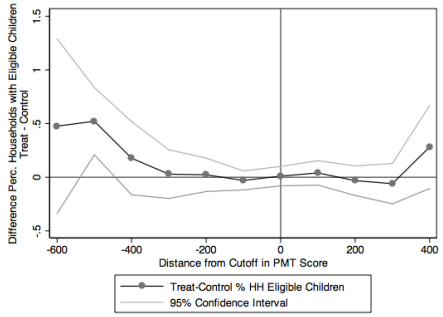
For asset ownership, the model predicts that the wealth effect will dominate the groups that are far below the cutoff, but the distortionary effect will dominate the difference between treatment and control just above the eligibility cutoff. If the utility cost of hiding assets is lower than the utility cost of not purchasing assets, reported ownership of hideable assets will be lower than reported ownership of non-hideable assets around the cutoff.

Although the coefficient on asset ownership is greater than zero for households below the eligibility cutoff and becomes negative for households with PMT scores 200 or more points above the eligibility cutoff, at no point is that impact different from zero. An extreme negative impact on hideable asset ownership for households with PMT scores 400 points above the eligibility cutoff is accompanied by very large standard errors.

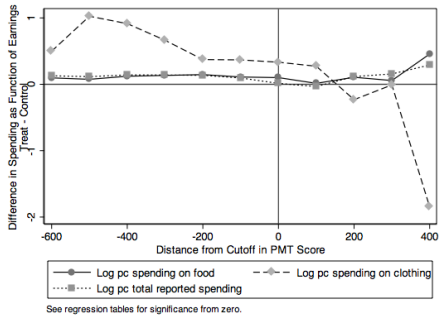
Figure 2.4: *Difference Between Treatment and Control in PMT-Applicable Characteristics, 1999*



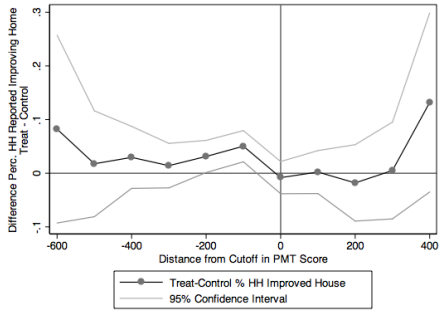
(a) Asset Ownership, Controlling for Baseline



(b) HH has Eligible Child



(c) Spending as Function of Earnings



(d) Household Improvements

The model predicts that the distortionary effect of the presence (or increase) of eligible-aged children in the home will not occur around the cutoff, because it does not make the PMT score more favorable for eligibility. Rather, households that meet eligibility PMT scores may be more likely to take eligible-aged children into their home in order to access or increase the size of their transfers. One would expect, then, to see an increase in the presence of eligible-aged children in homes with low PMT scores in the treatment group compared with the control group. If the children are transferred from wealthier family members within the same village, we may see a decrease in households with eligible-aged children in the treatment group among households with PMT scores above the eligibility cutoff.

There is a positive treatment effect of having children in the home at the poorest PMT scores, but the effect is only significant at one point in the analysis. I find no significant difference between the presence of eligible-aged children in the control and treatment groups at any other point along the eligibility spectrum. It is possible that households in treatment villages are pooling resources with households outside of the treatment village and inviting children into their home, but the evidence is not consistent.

If households decrease spending on assets, they will increase spending on other things. The model predicts finding that treatment households near the eligibility cutoff, particularly those with PMT scores just above the eligibility cutoff, will spend a greater portion of their reported income on items not included in the PMT score than control households. I look at spending as a function of earnings over three spending measures: food, clothing, and total reported spending, which includes a range of reported expenditure on items not included in the PMT score. I find no evidence of over-spending on non-PMT items. I find that the treatment group consistently spent more on these items than the control group, and that there is no discontinuity around eligibility.

Finally, I look at reported household improvement by distance from eligibility cutoff. This measurement is highly imperfect, because the variable does not specify whether the changes to the household were significant enough to change the households PMT score. As

with asset acquisition, we expect to find a wealth effect for households below the eligibility cutoff, and a distortionary effect for households just above the eligibility cutoff. Indeed, we do see that treatment households with PMT scores below the eligibility cutoff were, in fact, *more*, not *less* likely to make improvements to their housing structure than control households. This is particularly strong, and significant, for households just below the eligibility cutoff. The effect goes to zero at the cutoff, and dips below zero.

These findings suggest a wealth effect of positive impact on spending on household improvement, and a potentialy small distortionary spending effect for households near, but above, the cutoff.

The coefficients in a difference in difference that represents the graphs in Figure 2.4 are shown in Table 2.13, as identified through the analysis of Equations (2.11) and (2.12).

$$A_i = \alpha_3 + \gamma_1 \cdot T + \sum_{j=-7}^{j=5} \tau_j (C_r + 100j) \leq S_i \leq (C_r + 100(j+1)) + T_i \cdot \sum_{j=-7}^{j=5} \nu_j (C_r + 100j) \leq S_i \leq (C_r + 100(j+1)) + \epsilon_{3i} \quad (2.11)$$

$$A_i = \alpha_3 + +\gamma_{2a} \cdot P + \gamma_{2b} \cdot P \cdot T + \sum_{j=-7}^{j=5} \tau_j (C_r + 100j) \leq S_i \leq (C_r + 100(j+1)) + T_i \cdot P \sum_{j=-7}^{j=5} \nu_j (C_r + 100j) \leq S_i \leq (C_r + 100(j+1)) + \epsilon_{3i} \quad (2.12)$$

Equations (2.11) and (2.12) formalize the analysis shown visually in Figure 2.4. As with Figure 2.4, we propose that significant differences between spending among the control and treatment group will increase around the eligibility cutoff point for total and hideable assets. They may increase around and below the cutoff for eligible-age children as households increase their transfer size at any eligibility measure.

Table 2.13: *Differential Distortionary Effects by Probability of Eligibility Change: Assets and Children*

	(1)	(2)	(3)
	Total Assets	Hideable Assets	HH Has Eligible Children+
Post	-0.0289***	-0.0198	0.802***
	(0.007)	(0.013)	(0.051)
Post*Treat++	0.00635	0.00780	-0.0180
	(0.009)	(0.017)	(0.067)
Post*Treat*100 Below Cutoff	0.0108	0.0140	-0.00335
	(0.009)	(0.019)	(0.079)
Post*Treat*200 Below Cutoff	0.0122	0.0199	0.0318
	(0.010)	(0.020)	(0.134)
Post*Treat*300 Below Cutoff	0.00334	-0.0124	0.0756
	(0.016)	(0.031)	(0.195)
Post*Treat*400 Below Cutoff	0.0139	0.0190	0.328
	(0.015)	(0.030)	(0.258)
Post*Treat*500 Below Cutoff	0.0183	0.0446	0.557**
	(0.017)	(0.033)	(0.263)
Post*Treat*600 Below Cutoff	-0.0283	-0.0617	0.341
	(0.021)	(0.043)	(0.542)
Post*Treat*100 Above Cutoff	0.00601	-0.00797	0.00952
	(0.015)	(0.026)	(0.077)
Post*Treat*200 Above Cutoff	-0.0249	0.0223	-0.0825
	(0.023)	(0.043)	(0.108)
Post*Treat*300 Above Cutoff	-0.0402	-0.0402	0.0187
	(0.030)	(0.072)	(0.131)
Post*Treat*400 Above Cutoff	-0.0528	-0.166	0.353
	(0.044)	(0.105)	(0.221)
r2	0.0184	0.00955	0.541
N	45344	45344	45454

+ Dummy Variable for Whether HH has eligible Children

++Omitted Group has PMT score 0-100 points above eligibility cutoff

Note: Standard errors clustered at village level

* p <0.10, * *p <0.05, *** p <0.01

Table 2.14: *Differential Distortionary Effects by Probability of Eligibility Change: Spending*

	(1) Made HH Improvement	(2) Spending on Food+	(3) Spending on Clothing+	(4) Total Spending+
Treatment	-0.00817	0.101**	0.332**	0.0139
	(0.015)	(0.042)	(0.136)	(0.045)
Treat*100 Below++	0.0586***	0.00946	0.0351	0.0811*
	(0.015)	(0.041)	(0.107)	(0.043)
Treat*200 Below	0.0394**	0.0452	0.0402	0.121**
	(0.018)	(0.046)	(0.139)	(0.049)
Treat*300 Below	0.0223	0.0331	0.331*	0.126**
	(0.025)	(0.058)	(0.172)	(0.060)
Treat*400 Below	0.0377	0.0210	0.581**	0.128
	(0.033)	(0.088)	(0.246)	(0.079)
Treat*500 Below	0.0258	-0.0257	0.693**	0.101
	(0.053)	(0.115)	(0.278)	(0.106)
Treat*600 Below	0.0904	-0.00350	0.166	0.114
	(0.090)	(0.175)	(0.425)	(0.143)
Treat*100 Above	0.0104	-0.0863	-0.0602	-0.0478
	(0.022)	(0.057)	(0.162)	(0.059)
Treat*200 Above	-0.00982	0.00544	-0.564*	0.102
	(0.037)	(0.089)	(0.296)	(0.117)
Treat*300 Above	0.0133	-0.0440	-0.342	0.142
	(0.049)	(0.111)	(0.485)	(0.162)
Treat*400 Above	0.140*	0.356	-2.180**	0.277
	(0.085)	(0.374)	(0.955)	(0.485)
r2	0.00287	0.0844	0.0273	0.116
N	21052	17476	17476	17476

Omitted Group has PMT score 0-100 points above eligibility cutoff
+ Log Per Capita, regression includes controls for log pc reported earnings
++ *Below* and *Above* refer to relationsthip to eligibility cutoff where households below the cutoff are more likely to be eligible.
Note: Standard errors clustered at village level
* p <0.10, * *p <0.05, *** p <0.01

In this analysis, it is possible to examine how the difference between treatment and control changes at different PMT score groups. The findings are not notably different from the findings shown in Figure 2.4.

The wealth effect of Progresa eligibility is correlated with the distortionary effect. For this reason, it is useful to examine distortionary effects among households at different proximities to the eligibility cutoff: households well below the cutoff will demonstrate a pure wealth effect of the program, and households just above the cutoff will demonstrate a

distortionary effect of trying to have a lower PMT score than they would in the absence of the program. I find little evidence of either: treatment households just below the eligibility cutoff are more likely to make improvements in their home than control group households, and a subset of households in the treatment group are more likely to have eligible-aged children in the home. There is no significant effect on asset ownership or spending decisions.

2.6 Conclusion

In this paper, I develop a model for the distortionary effect of the dynamic use of proxy-means tests for determining eligibility for social programs. I propose that as the possibility of influencing eligibility increases, households will be more likely to take action to reduce their PMT score, by either selling, not purchasing, or hiding assets that appear in the PMT score. These distortions will cause households to spend more on other items, such as food or clothing.

I also identify the challenges that the wealth effect causes in identifying distortionary spending incentives caused by an income-increasing social program. I resolve this analytical challenge by assessing other evidence of distortionary spending: additional spending on non-PMT goods. I also assess signs of distortionary spending along the eligibility cutoff, where the incentives would be greatest.

I find inconclusive evidence that these distortionary practices occurred in Mexico's Progresa program. Households do not report reduced asset ownership or quality of in housing structure. In general, they report more spending on non-PMT items, but this is not restricted to households that would most benefit from distortionary spending. There is some evidence that among households poor enough to be eligible for the transfer, households are more likely to invite eligible-aged children into their home than control households at the same PMT score.

These findings may be unique to this particular implementation of the PMT score. Implementation was imperfect in that households with the same PMT score were assigned to a different eligibility status. There was poor documentation of the eligibility cutoff

above which no households should have been eligible and below which all households with children should have been eligible. Finally, there was some confusion over how eligibility was determined. Although this confusion appears to have been resolved before the follow-up surveys took place, it is possible that some households did not realize how their eligibility was determined, and thus did not take action to adjust their eligibility score.

The potential that households may take action to manipulate their PMT score in a dynamic setting exists, and merits additional research. Despite the high expense of regularly conducting the requisite household surveys, PMTs are particularly popular because they improve targeting. However, if they can be manipulated over time, the targeting benefits decline, and call into question the cost-efficacy of this method.

Chapter 3

Searching for the Devil in the Details: Learning about Program Design With Rugged Fitness Spaces[1]

3.1 Introduction

Development practice, and even more so academic development economics, has experienced a boom in "impact evaluations" using randomized control trials (RCT) methods. The method of randomly assigning units to "treatment" and "control" groups addresses the difficult, if not insoluble, problem of inferring causation from observational data. This methodological advantage has led to a veritable explosion in the use of RCTs in evaluating impact in ongoing or new programs by agencies, NGOs and by academics. Shah, Wang, Fraker and Gastfriend (Shah et al., 2013) estimate that there are now over 2,500 development RCTs. Vivalt (Vivalt, 2016) has a database with recorded results of over 600 evaluations, 80% of which are RCTs .

This RCT movement was premised on the notion that an RCT impact evaluation was a promising way to learn about development and provide evidence for what works. Now

[1]Co-authored with Lant Pritchett

is a good time to reexamine that premise. In particular, randomization as a technique is flexible and can be embedded into very different learning strategies than an "impact evaluation" that attempts to trace the impact through the entire causal chain (or "theory of change" or "log-frame") from inputs to outcomes by examining outcomes/impacts of populations exposed to particular program designs, through comparing treatment and control groups. The question is under which conditions the typical RCT-Impact Evaluation (RCT-IE) approach is an optimal, or even particularly fruitful, technique for learning about program efficacy.

We draw on our experience in attempting to use the standard RCT-IE approach to the design and implementation of a social enterprise of a job placement agency (JPA) to show that the usual approach fails if applied too early in the design process. This motivates a simulation analysis that compares two alternative learning strategies, the RCT-IE and a "crawl the design space" (CDS) approach (Pritchett, Samji and Hammer 2013) which provides more speed and flexibility in exploring program design options. The simulation shows that when the design space is even modestly high dimensional and the fitness function[2] even modestly rugged (both of which are controlled parametrically in the simulation) the RCT as a learning strategy to identify a program design produces outcomes that are both worse on average and more variable (e.g. sometimes the RCT approach performs well, sometimes very badly) than even a naive CDS learning strategy.

3.2 The Solution is the Problem?

One of us had an experience as a doctoral student that serves as a good example of the type of learning we propose in this paper. As a doctoral student, Sara Nadel (writing in the third person in this section) felt well-prepared with the ideal solo field research process:

1. Identify a problem. Be clear about what an ideal world looks like. Identify the anomaly

[2]We use the term "fitness function" rather than the perhaps more common in economic literature "response surface" as this paper was influenced by the idea of "tunably rugged" fitness spaces a la Kauffman (Kauffman and Levin, 1987) and by a recent simulation paper in the medical literature using "clinical fitness" as the outcome but nothing hinges on the terms.

in the market that prevents the ideal world from obtaining.

2. Write a causal model about the relationship between the problem and a market failure that may be causing that problem.

3. Identify an intervention that could resolve the identified market failure.

4. Review the existing literature to learn the evidence from previously-tested interventions.

5. Implement the program in a randomly-identified half of a target population, and compare the outcomes between the treatment and control populations. Improved outcomes in the treatment group suggest that the proposed market failure did exist, and the program becomes a proposed policy for resolving this problem moving forward.

6. Write a paper summarizing these findings.

7. If results are positive, expand and replicate this study elsewhere.

While the above process is the recommended process for dissertation-writing in her field, it turns out it is very similar to the recommended processes for intervention design and evaluation of development programs in multi-lateral organizations, government, and non-governmental organizations. The best approach for resolving problems is to identify the problem, write a model, review evidence about alternative interventions for resolving the problem, test, and, if successful, scale and replicate.

3.2.1 The Solution in Practice

Sara set about to implement this process.

Identify a Problem: Through her experiences living and working Peru, expanded upon by followup field visits during her studies, she identified a problem and a hypothesis about what caused it: Peruvian youth from marginalized households, despite rapid economic

growth in the country and their ongoing and expanding investments in higher education, were not securing formal sector jobs. At the same time firms complained that they struggled to find appropriate talent.

Write a Causal Model: She hypothesized that the signal of higher education as discussed by Michael Spence (Spence, 1973) had broken down during Peru's massive expansion of higher education offerings. According to the *Ministerio de Trabajo de Peru*, The number of people with higher education increased by 98% from 2001 - 2012, while the number of formal jobs increased by 38%. Talented youth from marginalized households had no effective mechanism to prove their skills and secure one of these increasingly competitive jobs even though they now possessed some higher education.

Spence Model

Spence (Spence, 1973) identifies two groups with differing marginal products both in work and education:

Table 3.1: *Spence Model*

Group	Marginal Product	Proportion of population	Cost of education level y
1	1	q_1	y
2	2	$1 - q_1$	$y/2$

- Group 1: Education is costlier compared with Group2 in terms of effort and productivity is lower.
- Group 2: Education requires half as much effort as Group 1, and productivity is twice as high.

In a world without a signal, an employer will presume that the productivity of each employee is the average of both, and pay accordingly:

$$q = q_1 * y_1 + (1 - q_1) * 2 * y_1 \tag{3.1}$$

However, employers may identify some optimal level of education, y^*, such that if $y < y^*$, the employer will know that productivity is 1 with probability 1, and if $y \geq y*$, the employer will know that productivity is 2 with probability 1. In this case, Group 1 will get $y = 0$, and Group 2 will get $y = y^*$.

Model of signaling and education in Peru

The hypothesis Nadel wished to study applied to how this model is corrupted in the following conditions:

- Education is granular, not on a spectrum. Individuals either have a college degree or not.
- Credit constraints limit access to college education.
- There is an increase in the number of providers of college education and a decrease in the quality (non-financial cost) of education. In Peru, the number of college-age people pursuing a higher degree increased by 98% between 2002 - 2012.

In this environment, individuals make the following optimization decision:

$$Max \frac{q(y)}{\delta} - y - c, \; s.t. \; c \leq C \tag{3.2}$$

where c is the cost of higher education, and C is the maximum cost that an individual can pay. In this environment, Nadel proposed to research the following hypothesis: *A reliable skill set signal $\rightarrow$ increased formal labor opportunities for people with $C < c$*. When $C \perp q$, the value of higher education as a signal of talent becomes nil.

Identify an Intervention: Identifying the model was more simple than identifying the intervention. Spence's model identifies higher education as the reliable skill set signal. If higher education no longer plays that role, the ideal intervention would offer a better skill set signal. Designing the intervention was more challenging than identifying the model. The applicability of existing research was useful only to a degree. Eventually, with the support of psychometricians experienced in our target population, she developed a test that would

evaluate skill sets and preference sets consistent with dedication to work. She was eager to test its applicability.

Review the evidence: While there is a growing body of research about the characteristics of young adults from low income backgrounds who succeed professionally (Rubinstein, Heckman et al., 2001), it was not clear that research focusing on the urban poor in the US would apply to the rural poor in Peru.

Table 3.2: *LogFrame of Skill-Set Signaling to Improve Job Placements*

Activities take place inside the organization			Outside organization	
Inputs →	**Activities** →	**Outputs** →	**Outcome** →	**Impact**
	skill set Test	Signal	Firms hire differently	Increased productivity Increased youth employment
	Applicants must take test		*Firms must use test results*	

Implement the Intervention: The intervention was to provide a more reliable signal about worker quality, allowing firms to hire higher-quality workers and pay them accordingly. However, the signal only becomes useful if the employer is able to attract high-quality workers. What the model above fails to consider is that the quality of jobs also varies, and job-seekers evaluate signals about the quality of a job when they choose where to apply, and they try to optimize over the probability of landing a job and the long-term rents of having that job. This led to another model, that of the optimization of an applicant.

Application: Encouraging applicant turnout

Assume that each worker expects to work in perpetuity upon securing a job. This is unrealistic given the high level of turnover, but does not change the equilibrium decision-making as long as the length of time that a job-seeker expects to work does not vary by job. Job-seekers apply to jobs sequentially. Everything is priced at 1 other than wage and revenues. All characteristics are unique for each job, j, and the hiring entity must adjust the perception of the job j to make it more favorable to potential applicants. Variables in decision-making regarding application to job j:

P_j - probability of successfully landing landing job j

W_j - Monthly wage at job j

e_j - Enjoyment of job j, which could include treatment of employees, cleanliness of facilities, likelihood of working late, etc.

S_j - Financial and time cost of applying to job j, including travel, printing resumes, childcare, etc.

$\bar{u}$ - outside option for predicted wage, time horizon and application costs of applicant. This could be an alternative job that the applicant has yet to secure, or a current job either outside the home or an informal family business.

The Job-seeker's optimization equations are:

$Max_j(P_j \cdot \frac{(W_j+e_j)}{\delta} - S_j)$, s.t

$P_j \cdot \frac{(W_j+e_j)}{\delta} - S_j \geq \bar{u}$

$P_j \cdot \frac{(W_j+e_j)}{\delta} - S_j \geq 0$

Given applicants' optimization process, the levers available to improve applicant turnout is to increase perceived P_j or e_j, or to decrease S_j. Iterations in the intervention tried to do all of these in addition to publicizing the job opportunity to more people in order to find more people for which the optimization equation support applying for the position in question. Table 3.3 reviews adjustments made during the launch of this evaluation mechanism with the goal of receiving more (high-quality) applicants and encouraging those applicants to complete the application process.

Small adjustments that were designed to lower the cost of applying by asking less of an applicant (planning ahead is a big non-financial cost for our talent pool) such as the amount of time between calling the applicant and the date of their test or interview or receiving applications by text message had big effects on turnout. Using the name of the client instead of a generic hiring announcement in the job publicity increased the P_j or e_j because the job was perceived as a serious opportunity because the employers had better name recognition.

Table 3.3: *Learning at our Job-Placement Program*

City Date of request	**Piura** Aug 2012	**Piura** Oct 2012	**Piura** Feb 2013	**Piura** Mar 2013	**Chiclayo** Mar 2013	**Chimbote** Mar 2013	**Arequipa** Mar 2013	**Huanuco** May 2013
Announcement Mechanisms								
Newspaper	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Computrabajo*	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Bumeran*	No	NT	No	No	No	No	No	No
Facebook Page	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Paid Facebook Ad	No	No	No	No	Yes	Yes	No	Yes
Flyer	No	NT	No	No	Yes	Yes	No	Yes
University Career Counseling Centers	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Other Operations								
Used Company Name (instead of Farolito)	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Invitation to test sent $<$ 24 hrs. beforehand	No	No	No	No	Yes	Yes	Yes	Yes
SMS reminder 12 hours before Test	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
SMS applications accepted	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Schedule test online?	Yes	Yes	No	No	No	No	No	No
Success Measures								
Position viewings	NT	NT	783	128	236	20	119	NT
Filled out first filter	NT	NT	539	97	121	20	54	243
Met basic requirements	NT	NT	418	54	72	12	27	188
Finished Application	NT	NT	372	52	63	12	21	188
Took Test Invited/showed up	NT	NT	197	31	43	NT	18	152
Recommended (when based on test)	NT	NT	61	NT	11	NT	4	NT

*Common online job boards in Latin America
NT: Not Tracked

Other adjustments such as paying for Facebook advertisements, improved the applicant turnout and the quality of the candidates ultimately recommended for the job in some cases but not in others. For example, the position in Chimbote, despite the extensive publicity, generated a total of 20 applicants. However, the same combination of publicity turned out great candidates in Huanuco. This differential effect could be related to the population in each city (preferences and professional alternatives), thus, population preferences and outside options are a dimension that should be considered in the design space.

This type of learning was crucial to the the intervention, although it had nothing to do with the original model that Nadel built or the problem she aimed to solve. If she had written a paper about the impact of providing better skill set signals to matching in the labor market, this background research would not have been included. Had the paper concluded that there is an opportunity to improve matching in the labor market through better skill set signaling without mentioning all of this background learning, practitioners seeking to replicate her success would begin having to rerun this learning process over again. Alternatively, had she not engaged in this learning process at all, she would have found limited impact of better signaling because she would have not had enough high-quality applicants for her partner employers to filter acceptably. Efforts to attract users on both sides highlighted two characteristics about real-world implementation that her model was not prepared to incorporate:

- *Granularity and High Dimensionality* In trying to encourage users to adopt the test, it became necessary to revise aspects of the program that would be considered "insignificant" in the research context. Small iterations involved designing a better logo, creating a fancy information sheet, updating the intervention company website, and others. But while it was easy to see how "reliability of the signal" fit into the causal model, it was harder to see where specific adjustments fit into a model. How could she quantify a better logo or changing the logo color from blue to orange? The granularity of the interventions complicated their role in her model.

- *Ruggedness* Adjusting small characteristics of the program highlighted the number of

small adjustments that can generate big changes in outcomes. There were hundreds of things to consider in an actual program, and the combination of those things generated drastically different responses as well.

The importance of such small characteristics that were seemingly irrelevant and hence not included in discussions either of the model or in the "evidence" but highly consequential to the implementation highlighted holes in the *construct validity* of the approach. That is, models and evidence often presumed that useful discussions could be had about *classes* of programs/projects/policies roughly independent of the consideration of the granularity in design that distinguished *instances* of the class. That is, one could consider theoretically and empirically the *class* of "skill set signaling job placement programs" along the relative few dimensions identified in the theory (e.g. the reliability of the signal) and the relatively few instances of the class empirically examined. This experience with a failure of construct validity about classes of programs due to dimensionality and ruggedness was not unique, it is increasingly discussed in a variety of domains.

For instance, one might imagine that one could usefully talk about the theory and evidence of the impact of the *class* of programs that increased the availability of textbooks on child learning in school. As Muralidharan (Muralidharan, 2015) describes, there are now four well-identified estimates of the impact of providing additional textbooks in low income, low input availability settings. All four studies show zero impact of the provision of additional textbooks on the learning performance of the typical child. A mechanically implemented "systematic review" of the "high quality" evidence might justifiably conclude that there was no evidence that textbook availability mattered for child learning. But, as Muralidharan emphasizes each of the studies point to elements of the design space that interact with the impact of textbooks. For instance, in (Glewwe, Kremer and Moulin, 2009) the authors conclude the limited impact on the average fourth grade student is because the textbooks were too hard and hence only benefited the top students.[3] Other studies

[3]As our paper is about the dynamics of learning from various methods it is worth pointing out this Kenya study was initiated in 1995, with textbooks provided in 1996 and 1997. The NBER version of the paper appeared

have other *ex post* evidence supported explanations of the lack of impact: teachers didn't actually open the textbooks but stored them (Sabarwal, Evans and Marshak, 2014); when the provision of textbooks was anticipated, households did not purchase them (Das et al., 2011); or the textbooks only had positive impact when interacted with teacher performance pay (Mbiti, Muralidharan and Schipper, 2015). Thus we see the the class of programs called "provision of textbooks" has a high dimensional design space, elements of the design interact with availability, the fitness function is rugged over that design space, and *ex ante* program design often misses many of the elements that determine performance.

In a world where the design space is high dimensional and where seemingly small design changes can have big impacts on outcomes, the step from *writing a causal model about the relationship between an outcome and the market failure that causes it* and *implementing a program to resolve that market failure based on existing evidence* is a much more complex process than the recommended steps for program design currently recognized. Hence the desire to build into program design a mechanism for learning about program efficacy.

3.3 Simulating the Performance of Alternative Learning Strategies

We are going to address the performance of different learning strategies in cases like our JPA in Peru in which we consider that the fitness function is rugged over a high dimensional design space. We do this by building an artificial world that abstractly represents some key features of the learning problem, and examining the results in this artificial world. The advantages of simulation are that the "truth" is known – in a way never possible in a real world of human beings – and that we can parametrically alter the world to examine how it affects the relative performance of alternative learning strategies.

The simulation contrasts the performance of two alternative approaches to learning.

---

in 2007 and the published version in 2009. While the delay in this case is perhaps an extreme example, it is not unusual for such learning to be made public long after program implementation.

The RCT learning strategy:

- Starts at a random point in the design space.
- Implements that chosen program design as the base case and additional local alternatives (treatment arms) over relatively long periods (say a ”year”).
- At the end of a year does statistical calculations and moves from the base case to the alternative if the alternative is statistically significantly better.
- In the second year the base case and a local alternative in a different direction are evaluated and again the RCT moves to the new program design if the treatment arm is statistically significantly better than the treatment arm.
- The result of the RCT learning is the program design at the end of the two year period.

The CDS (“crawl design space”) learning strategy:

- Starts at the same random point in the design space as the RCT strategy.
- Implements that program design and one other chosen randomly (with replacement) from the entire design space.
- At the end of one “month” compares the outcomes and moves to the best of the two, with no account of statistical significance.
- Repeats this procedure for 24 months (two years).
- The result of the CDS learning is the program design chosen at the end of the two year period.

We can compare at the end of the two-year period the performance of the RCT and CDS program designs to each other and to the best possible program design for a given fitness function.

3.3.1 Simulation: Design Space and Fitness Function for Farolito

The Job Placement Agency (JPA) objective is to find people who are a two-sided match to the JPA's contracting firm's available jobs. Each person has a job specific productivity or aptness for the particular job and one element of match is that the productivity is high enough that the firm wants to retain the person. Each person also has a hedonic match to the particular job such that, if hired, he or she would choose to stay on the job. We define success for the JPA as placing people with the contracting firm who are above a threshold in aptness and hedonic and hence are a two-sided match.

We assume there is a flow of P people in each period (call it a "month")[4] who are potentially interested in a job. The JPA program design problem is to attract people to its services and then identify two sided matches.

In order to be able to visualize the fitness function over the design space as a 3D graph we limit the JPA program design space to two dimensions: a communications (C) strategy to attract applicants from P possible people and a filter (F) element that creates the signal of aptness for an employer. Each of C and F have N possible options and hence the design space has N squared elements.

Communications Strategy

This is simulated as a draw of NP people from a random distribution on aptness and a random draw on hedonics where we can control the correlation coefficient of aptness and hedonics.

C represents how we communicate with applicants in order to attract applicants. For instance, the variables that we adjusted, in order from lowest to highest density in terms of price, are:

1. Email & online only

[4]We say "month" and "year" in quotes as these are just arbitrary periods and any mapping from elements of the software code to elements of the world is allegorical but, to avoid pedantry, we will use terms like month and year to describe our artificial world without scare quotes each time.

2. Receive applications online, communicate by email and text message

3. Email, Text message, plus one phone call to invite to the test

4. Email, Text message, phone call to invite to the test, plus a reminder text message

We represent a C design as a triplet: (c_1, c_2, c_3).

A person j from the NP people applies to the JPA if:

$C_j = c_1 + c_2 * a_j + c_3 * h_j + \epsilon_{c,j} > C_{threshold}$

This is simple and intuitive. Communications strategies can either try to attract more people to apply (an increase in c_1) or try to induce high-ability people to apply (an increase in c_2) or try to induce people with a good hedonic match for the job to apply (an increase in c_3). It is obvious that there is a trade-off of different types of errors. Suppose the communications strategy, in a communications attempt to attract only high quality applicants discouraged people whose aptness was in fact above the threshold. Then there are potential successes who are never seen and their exclusion from the interview process reduces the total number of successful placements. Conversely, if the communications strategy attracts many on the basis of hedonic match who do not meet the requirements, then for a given filter applied by the JPA, more "bad hires" would be made: people hired but were a mistake because they were not in fact highly apt or productive.

The application of the C strategy results in some proportion of the NP pool of people applying to the JPA.

Filter Strategy

F strategy represents how we filter applicants. Computerization adds a level of difficulty among our job applicant base. As such, in order from lowest to highest density, the variables are:

1. Group interview

2. Handwritten test

3. Online filter which confirms that the applicant meets basic job requirements, completed at home

4. Computerized test given in a supervised environment

The second element of the JPA program design is the application of the filter, which is two sided. That is, the application of some assessment produces an estimate of the aptness and hence filters out as hires those below the estimated aptness threshold but it is also the case that applicants may choose to drop out of the recruitment process as a result of the experience of the filter. Hence the F strategy is (f_1, f_2).

A person is considered hired if:

$A_j = f_1 * a_j + \epsilon_{a,j} > A_{threshold}$ (the person is estimated by the filter to be above the critical threshold on aptness)

and

$H_j = f_2 * h_j + \epsilon_{h,j} > H_{threshold}$ (the person, even after exposure to the JPA filter process, wants the job).

The application of the filter F to the applicants produces some number (perhaps zero) of the P pool of possible applicants in a given month who are hired.

The number of successes in a given month is the number of hires, based on their actual aptness and hedonic match with the job, who are truly above the threshold as this implies they will be hires for whom both the firms desire them to stay and who desire to stay. Since the JPA client firm wants to fill the position with productive hires and low hiring cost per applicant, this is the desired outcome.

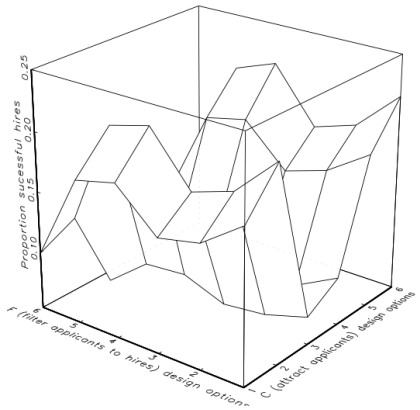
Fitness Function

We create our artificial world to have a fitness function over the design space which is *rugged* in a manner that is parametrically controlled[5]. The ruggedness has three elements

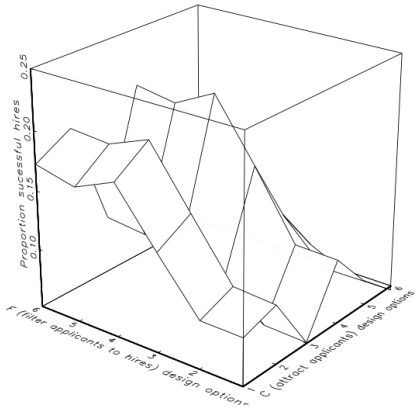
[5]There is a literature, initiated by Kaufmann (with others) on NK models which provide a "tuneably rugged" fitness landscape by altering N and K. We chose an alternative approach as the NK model has less verisimilitude for the JPA problem.

illustrated in the examples in Figure 1. First, the fitness function is not linear (or quadratic) in one strategy conditional on the other. Second, the fitness function is interactive, the relative outcome of F_j versus F_k for C_j is not (necessarily) the same as for C_k. Third, these outcomes differences across (C,F) strategies can be parametrically made "big" or "small" in the relevant fitness metric even across "local" alternatives. There is an optimal strategy of communications and filter (C^*, F^*) but proximity in the design space to the (C^*, F^*) combination does not ensure proximity to the optimal outcome.

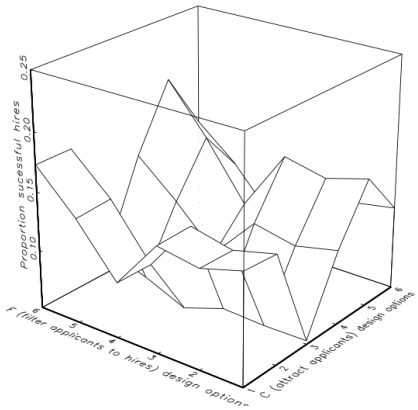
Figure 3.1: *Three Examples of Rugged Fitness Function*



(a) Fitness Function Example 1



(b) Fitness Function Example 2



(c) Fitness Function Example 3

A fitness function for any given "context" (where "context" can be place, implementing agency, time-varying job market conditions, etc) is determined by fixing the parameters of the communications and filter strategies.

For C the first strategy was always $C_1 = (0.5, 0, 0)$ which was the "default" strategy that attracted applicants uncorrelated with either aptness ($c_2 = 0$) or match ($c_3 = 0$). For the remaining strategies the elements were chosen randomly from the possibilities: $\{$ -0.5, 0, 0.2, 0.5,1$\}$. For example, a C design triplet (0.2, -0.5, 1) would attract fewer default applicants than the base strategy but attract applicants in inverse correlation with aptness but positively selected on hedonic match. This 3 by N matrix was then inflated or deflated by a ruggedness parameter.

The elements of an F design were chosen from the possible values 0, 0.5,1,1.5 or 2 randomly for each element for each of the N strategies for C and hence the F parameters are different for each C strategy. For instance, an F design of (1.5, 0) implies the filter process would strongly select on aptness and dropout would be uncorrelated with match at the filter stage.

These choices fix the fitness space as they determine who is attracted to apply for the job via the communications strategy, who is offered the job as a result of the filter and who takes the job as a result of hedonic match.

Success is determined by whether those who are offered and take the job are *in fact* good job matches. That is, our measure of JPA success is the fraction of the population who are place and are truly above the aptness and hedonic match thresholds. That is, the filter on aptness can be "good" or "bad" in selecting on true aptness. Communications and filter strategies have differential performance because they can make mistakes of various types: they can attract too few applicants but successfully filter those who do apply but result in too few matches relative to the optimal; they can attract lots of applicants but of the wrong sort (e.g. the communications process can attract ill-suited applicants on aptness or hedonic match); and then, for a given filter too many of the wrong people will not be hired.

3.3.2 Simulation: Learning Strategies in the Artificial World

We have built this simulation to examine questions about learning. Suppose we were facing a rugged fitness function over a high dimensional design where the fitness function is contextual (or at the very least we do not know *ex ante* if it is contextual or not). That is, the efficacy of a particular (*C*,*F*) design might be different in Detroit versus New Delhi versus Cairo and hence cannot be assumed to apply to Peru (or to different cities in Peru). What is an appropriate learning strategy as a sequence of actions and feedback loops from those actions that would be likely to lead over a fixed period to a good program design, a combination of *C* and *F* that produces a high (of not optimal) outcome?

We assume the fitness function over the design space is fixed over the period of the simulation but unknown for a "context" (where context can include country, region, implementing organization, availability of other alternatives, etc.). We are going to simulate a period of 24 periods (months) with observed feedback on outcomes at the end of each month[6]and apply two different learning strategies (CDS and RCT). Each of the two learning strategies starts at the same point in the design space and then, relying on the feedback from outcomes, dynamically alters the (C, F) strategy being pursued. We then compare the strategies at the end of the period to see which was better at learning, on average, over a variety of possible fitness spaces.

Learning Strategy: CDS (Crawl Design Space)

Both CDS and RCT start at a given program design, (C_0, F_0) (where "0" indexes time not strategy number).

Two strategies are implemented in the CDS learning strategy: the current best and one alternative. The alternative is chosen each period from all other strategies besides the current best. Sampling on alternative strategies is with replacement so previously-tried

[6]This involves a modest elision on the actual dynamics as we assume at the end of each month we observe not just the actual placements made by the (*C*,*F*) strategy but also the *successful* placements. This essentially assumes that the "true" aptness and hedonic match are revealed instantaneously after the job has begun. In reality it would take some time for this to be revealed.

strategies can be tried again. At the end of each period t the outcomes of successful hires from the two strategies are compared as simple averages (same as counts as the number of potentials is fixed at NP) and:

If:

$FF(C_{CurrentBest,t}, F_{CurrentBest,t} > FF(C_{Alternative,t}, F_{Alternative,t})$

then the current best program is retained; and:

$(C_{t+1}, F_{t+1}) = (C_{CurrentBest,t}, F_{CurrentBest,t})$

if the alternative is better then it is adopted; and:

$(C_{t+1}, F_{t+1}) = (C_{Alternative,t}, F_{Alternative,t})$

At the end of the periods (24 in this simulation exercise) the resulting program design is:

$(C_{CDS,T}, F_{CDS,T})$

and hence the outcome of implementing that program design is:

$FF(C_{CDS,T}, F_{CDS,T})$.

We acknowledge this is an extremely simplistic learning strategy. There is no optimal choosing of the starting point, no attempting to guess what a good "alternative" strategy might be, no use of "statistical significance" to choose whether to switch, no memory to the learning process (e.g. if a program design is outperformed in one month it is replaced even if it has worked for months against other alternatives). We intend this to be a cave man simple search/learning strategy as we are not searching for the "optimal" learning strategy[7] rather we are attempting to articulate a simple learning strategy that pretty much any organization could implement.

Learning Strategy: RCT

The learning strategy for RCT in this artificial world is intended to mimic the standard RCT which chooses a "treatment" and perhaps a few alternative "treatment arms" and then holds the treatment fixed for a period sufficient to generate statistically significant results

[7]We are reasonably confident from the large literature on optimization there is no generally "optimal" algorithm for finding the optimum of an arbitrarily rugged fitness function.

and hence shifts treatments relatively infrequently.

We model this by having the RCT start from exactly the same starting point as CDS. This default or "current best" program design is tested against one other alternative, which is a local alternative[8] that alters the strategy in just one dimension. In our simulation we first search in the C dimension and then in the F dimension.

The principal difference is that, in the interests of "statistical power" and to maintain the "integrity of the experiment" rather than making modifications to the program design each month, the program is changed only at the end of one year (12 periods/months).

At the end of 12 periods, the "current best" (which is the initial at the end of period 12) is compared to the alternative. The alternative is adopted only if it is statistically significantly better than the "current best." Then, having adopted the new strategy for period $t + 13$ and on the alternative is chosen as a local alternative in the F dimension.

Then, at the end of 12 more periods (and hence the end of the first two years of implementation) the current best is compared to the alternative and the alternative is adopted if it is statistically significantly better than the current best.

The result is an RCT strategy and outcome:

$(C_{RCT,T}, F_{RCT,T})$

and hence the outcome of implementing that program design is:

$FF(C_{RCT,T}, F_{RCT,T})$.

The simulation is built around the following key differences between the learning strategies:

1. CDS learning updates the default program design upon any superior outcome, while RCT learning updates only on a statistically superior outcome.

2. CDS is faster, updating once a month compared to RCT learning which is once a year. By extension, the RCT measurements are more precise, with lower variance and less

[8]Since we do not assume the design space alternatives are ordered in any way we treat "locality" as a cycle so that alternative 1 is "local" to both alternative 2 and alternative N, where N is the number of possible choices for either C or F designs.

likely to be influenced by noise.

3. CDS learning chooses a program design variant from the universe of strategies. RCT learning only adjusts one element, first C then F.

While one might object that this simulation is "cooked" in favor of the CDS over the RCT strategy our response is three-fold. First, we think that the description of program design changing at best once a year is not a complete caricature of the actual practice of RCTs. We personally have been acquainted and/or directly involved with a several RCT studies in which a treatment arm was obviously badly failing but the experimenters needed to maintain it to fulfill the study design.[9] Second, we also feel it is not a terrible caricature of RCTs to imagine one "main" and one "alternative" treatment arms. The interests of statistical power with potentially noisy measurement and modest impact size tend to limit the number of treatment arms to a small integer. Many just implement one program and test it against the counterfactual of "no treatment," others have one alternative, some (but few) as many as four. Third, given the attention that "impact evaluation" style RCTs have received as a method for learning about "what works" in development, if it is really so easy to "cook the books" against them as a learning tool this is in itself revealing.

3.3.3 Mechanics of the Simulation

The basic structure of the simulation is:

Step I: A fitness function is created as a random choice over the possibilities for the five parameters (c_1, c_2, c_3) and (f_1, f_2)) and a ruggedness parameter.

Step II: Each month a universe of NP (where NP is 1000) individuals are exposed to the C strategy and choose whether to apply to the JPA. The JPA applies some set of actions after which the JPA filters candidates and candidates choose whether to take the job.

[9]One of my (Lant Pritchett) first trips to an RCT that was a collaboration of an NGO and an academic partner the head of the NGO introduced me to junior worker from the RCT partner saying "This is Dan, his job is to make sure we don't help any children" as the academic partner was encouraging them to stick to a treatment arm the NGO had recognised as flawed and wished to abandon.

Step III: The "true" results of number of successful hires at the end of each month are known.

Step IV: Based on the results the (C,F) strategy is updated according to the rules of CDS (monthly) or RCT (yearly).

Step V: At the end of the T periods (where T is 24 months) the results of the final strategy of the two strategies are compared at the overall fitness function (that is, the success of $(C_{CDS,T}, F_{CDS,T})$ and $(C_{RCT,T}, F_{RCT,T})$ are compared when applied over data for all 24 periods as an approximation of the "true" results as the monthly results depend on the random variances in the various selection functions).

Step VI: After this process is repeated for I iterations then the results for having applied the CDS and RCT strategies over a large number of different fitness functions of given ruggedness is known and the average and variance of outcomes of the strategies can be computed.

3.4 Results of the Simulation

The purpose of the simulation exercise is to ask: "In this artificial world built to capture challenges of designing a program when facing a high dimensional design space and rugged fitness function what are the implications of various learning strategies?" The results show the RCT strategy is a low mean (the learning gain is smaller) and high variance (the risk of getting really poor results is larger) learning strategy compared to CDS.

3.4.1 Baseline Results

These simulations produce two main results, illustrated in Table 3.4 using design spaces from 5 to 10 options for each of C and F (between 25 and 100 possible program designs).

First, the CDS learning strategy typically reaches a substantially better program design than the RCT learning strategy.

The second column of Table 3.4 shows the average over 1000 iterations on different fitness functions and different starting points of the excess performance of CDS over RCT

scaled as a function of the gap between the "best" and the "average" for each fitness function. Since the absolute numbers are more or less arbitrary (e.g. we can produce more or less average success by varying the threshold cutoff of the filter or hedonics) we feel this is a natural metric for the gain from learning: *how much of the distance between just having picked a strategy at random (which would produce the average result) and having reached the best possible result was closed by the learning one did?* For 6 options (which is our default), the superiority of CDS over RCT is 49 percent of the best versus average gap. Interestingly, the gain of CDS over RCT actually declines with number of options as the CDS is able to evaluate a smaller fraction of the total design space. At 10 options (design space of 100 possible programs), the CDS over RCT gain is still 47 percent of the best to average gap.

To illustrate the intuition with absolute numbers, in the baseline parameters and 6 design options for each element the average success is 10.2%. The best result in each fitness function averaged over 1000 fitness functions is 15.7% successful hires. The average CDS result is 14.9% so nearly reaches the best result and makes substantial improvement over its starting point which, since it is randomly chosen, is the average result (with six variables it is not surprising as up to 24 options are evaluated). The gain from CDS learning is improving success from 10.2% to 14.9%. The average success rate for the RCT learning strategy is 12.2%. Hence the gain of CDS over RCT is: (14.9-12.2)/(15.7-10.2)=0.49.

Interestingly, the learning gain of CDS relative to RCT gets somewhat smaller as the number of options (hence the dimensionality of the design space) gets larger. This is because the RCT strategy does about the same in gain but the gain of the CDS as a proportion of the possible gets lower as the ratio of the 24 trials to the total design space gets smaller.

Table 3.4: *Simulation Results*

(1) Number of options	(2) Gain CDS over RCT to max over average possible*	(3) Percent excess of RCT standard deviation**
5	0.516	1.57
6	0.490	1.64
7	0.487	1.65
10	0.467	1.74

*$((C_{CDS,T}, F_{CDS,T}) - (C_{RCT,T}, F_{RCT,T}))/(BestPossibleSuccessfulHires - MeanSuccessfulHires)$
**(Standard Deviation of RCT Results / Standard Deviation of CDS Results)

The second main result is that the CDS learning strategy has a lower variance across alternative fitness functions than does the RCT strategy. Column 3 of Table 3.4 shows that the ratio of the RCT to CDS standard deviation is 1.5 times higher with a design space of 25 options and 1.74 times higher when there are 100 options.

The intuition in the simulation is as follows: the RCT has fewer moves across the surface of the fitness function and so if, by random chance, happens to get started with a bad program design and this happens to be in a bad neighborhood of the fitness function, it could end up with a very poor outcome. This is not always the case, because ruggedness means that a bad program design could be close to a good one. Thie plays out in our simulation. In a run of 1000 simulations of the baseline parameters and six options, the 10th percentile result for the RCT was 0.106–which is substantially worse than the average outcome of 0.14.

3.4.2 Performance of Learning Strategies across Degrees of Ruggedness

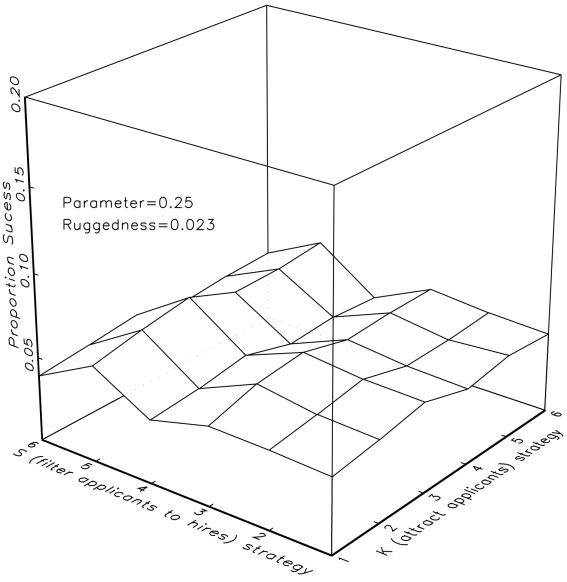
In our artificial world, we can parametrically alter the ruggedness of the fitness function by scaling up or down the coefficients that determine performance. Figure 3.2 shows typical fitness spaces when, relative to the baseline ruggedness parameter of 1 (which produces the graphs in Figure 3.1) the ruggedness parameter is either 0.25 (producing the smoother surface) or 4 (producing a more rugged surface). We can compute the actual ruggedness as

the average absolute deviation of fitness at each element of the design space compared to its eight local neighbors.[10]

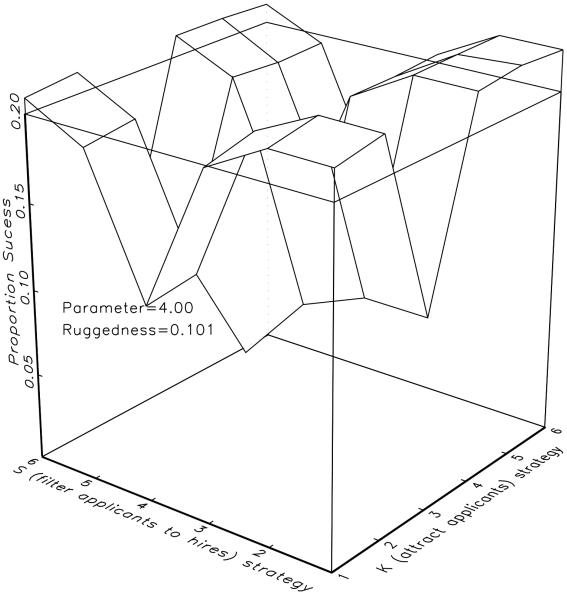
Table 3.5 shows the results of holding all our parameters as in Table 3.4 for six options and then varying only the ruggedness parameter. As designed, from the lowest to highest values the ruggedness increases by a factor of five from .020 to .103. The striking, even if expected, result is that the ratio of the standard deviation of the RCT strategy to the CDS strategy increases from roughly 1 (they do about the same) to 4.25. Practically (if a simulation result can be such) this says: if the fitness function is very rugged then design matters a great deal and hence the losses from not crawling the design space can be vary large and the results of the RCT strategy can be good (if one happens to start at or near a good alternative) or very bad (if one happens to start in a bad neighborhood).

[10]This again assumes that since the design space isn't ordered the space "wraps around," i.e. graphically the local neighbors can be computed by tiling the design spaces side by side. Hence the eight local neighbors of design space point (1,1) are (clockwise from northwest): (6,6),(6,1),(6,2),(1,2), (2,2), (2,1),(2,6),(1,6).

Figure 3.2: *Comparing smoother and more rugged fitness functions*



(a) Smooth Fitness Space



(b) Rugged Fitness Space

Table 3.5: *Learning Results Varied Across Ruggedness of the Fitness Space*

(1) Ruggedness parameter	(2) Ruggedness (absolute difference)	(3) Gain CDS over RCT (ratio to max less average)	(4) Percent excess of RCT over CDS standard deviation
0.25	0.020	0.319	1.04
0.5	0.042	0.445	1.19
1 (base case)	0.074	0.489	1.64
2	0.094	0.461	2.36
4	0 .103	0.412	4.25

Interestingly, the average gain of CDS over RCT (as a ratio of best to average) first grows with ruggedness and then declines.

3.4.3 Other Variations on the Base Case

The results presented in Table 3.4 just varied the simulation across the number of program design options but kept all other parameters of the simulation constant. It is possible that the two main results are fragile with respect to minor variants in the artificial world. In this section we vary three elements of the simulation to examine the robustness.

Correlation of aptness and hedonic fit. In the base case we assumed that there was no correlation between a person's aptness in a job and their preferences to be in the job. Since the program designs often affect the two characteristics of the pool of people differently (e.g. communications strategies may attract the apt but deter well-matched or vice versa), this could affect the results. In Table 3.6, the first row are the results for the base case parameters with 6 options for each element of program design (36 total designs). In the second row the correlation of aptness and match was increased to 0.8. This produced roughly similar results with modestly higher average success and even more difference in the RDS and CDS variation.

More noise in decision rules. One feature of the simulation is the extent to which the program design versus randomness (including the inability of program design to effectively target the "right" applicants through communications or filter the most apt applicants with an instrument) affects outcomes. We add more noise, the individual specific ϵ's in the various equations, to the process. The result is that the learning advantage of the CDS over RCT persists but is lower and the relative variability of RCT versus CDS final program design outcomes declines. Again, this is intuitive as it increases the uncertainty of identifying precisely the program design versus random elements month to month and hence reduces the ability of the CDS to identify a good program design as it is likely to wander away from a good design due to noise.

Table 3.6: *Variations on the Base Case*

(1) Number of options	(2) Gain CDS over RCT to max over average* possible	(3) Percent excess of RCT standard deviation**	(4) Average Success***	Description of parameter changes
6	0.490	1.64	0.102	Base case parameters
6	0.504	1.84	0.163	Correlation of aptness and match 0.8 (instead of zero)
6	0.371	1.07	0.057	Variance of the noise in decision rules increased
6	0.457	4.99	0.025	Higher threshold in ability and hedonic match for success

*$((C_{CDS,T}, F_{CDS,T}) - (C_{RCT,T}, F_{RCT,T}))/(BestPossibleSuccessfulHires - MeanSuccessfulHires)$
**(Standard Deviation of RCT Results / Standard Deviation of CDS Results)
***(% successful hires from a population of 1000

Higher thresholds of success. In the base case, the thresholds for aptness and hedonic match ($A_{threshold}$ and $H_{threshold}$) that define successful placement were set to zero. If we increase that to 1 this reduces the success rate from .102 to .025 on average, which is intuitive. This produces roughly the same learning gain from CDS over RCT but also dramatically increases the RCT variability as it increases the gap between successful and failing program designs which increases the risk the RCT ends at a relatively low performance final design.

3.5 Is the Fitness Function for Social Programs Rugged? Evidence about what "Evidence" Means

We show that in an artificial world in which the design space is even moderately high dimensional and the performance is sensitive to program design (rugged fitness function) a fast turnaround learning procedure dominates a more precise and statistically reliable but slower and more local learning process in both mean gain and in the variance of expected outcomes across contexts. This finding about an artificial world is obviously of little interest unless it captures features of the "real" world of development program/policy/projects.

The next two sections argue that: (a) the available evidence from the cumulative RCTs suggest that there the well-known issues of *external validity* (that the whole fitness function may vary across contexts). There are also major issues of *construct validity* in the *classes* of policy/program/projects where there are attempts to summarize what the evidence says about "what works" are insufficiently granular to be of use to practitioners and (b) to the extent that the new concerns about "behavioral" economics are important construct validity in development projects/policies/programs becomes nearly impossible.

3.5.1 Heterogeneity in Estimated Impacts: External Validity and Construct Validity

We found that in the RCT learning strategy, variance is increasing in the ruggedness of the fitness function highlights the role of "construct" validity in assessing evidence. We

give examples of the sensitivity of outcomes to program design from a number of domains in which RCTs have been popular, and then discuss the findings of reviews of RCTs. As one looks across nearly all domains in which there have been sufficient evaluations (e.g. microfinance, training, education, income generation/livelihoods, community development, etc.) one finds large differences in the estimated impacts. While these differences are hard to disentangle, many appear to be construct validity and not purely external validity issues due to "context"[11].

3.5.2 Examples of Program Ruggedness from Impact Evaluations

Impacts on learning McEwan (McEwan, 2015) reviews the rigorous evidence about impacts on learning in primary schools of developing countries and compares the average impact across 11 classes of interventions like "Computers or technology" or "Instructional materials" or "Deworming drugs." The average effect size (gains as ratio to student standard deviation) 0.072, and the standard deviation across these 11 classes of interventions is 0.05 and the range is 0.161 (from 0.15 to -0.011) and the highest average impact size is for "computers or technology" with an average effect size of 0.15. However, if one looks across the 32 "computer or technology" interventions the range is more than 1 full standard deviation from positive 0.45 to *negative* 0.58: the "within class" range of instances within the class is six times the range across averages of classes of programs. While this might be chalked up to "external validity" or cross-contextual heterogeneity in impacts, some of the largest differences are across treatment arms in the same context. So for instance, the same study in India found that "after school" computer-assisted instruction had an effect size near 0.30 whereas "in school" computer assisted instruction *reduced* learning by 0.58 standard deviations–an across treatment arm range of 0.88–more than 5 times the range across all classes of interventions. This suggests a very rugged fitness function for "computer or technology" learning interventions and makes a comparison of mean effects of "computers

[11]Where "context" is itself under-specified as people think of "country" or "place" as a catch-all for "context" but "context" could well be (and has been shown to be in some applications) regional, organizational, personal, path-dependence on history, existing alternatives to suppliers, etc.).

or technology" at 0.15 versus "contract or volunteer teachers" at 0.101 seems inconsequential relative to the within class impacts of differential design.

Evans and Popova (Evans and Popova, 2015) review six "systematic reviews" of the literature on how to improve learning in basic education in developing countries and show that the "systematic reviews" of the "rigorous evidence," even of the same topic (and two were by the same organization), often come to very different conclusions. This is not surprising when the within-class of intervention variance is high relative to the across-class mean differences and in that case small changes in the methods and filters to search for and include studies can lead to different results as the heterogeneity in impact estimates is so large across studies.

The high dimensionality and ruggedness of the fitness function is illustrated by the finding in Glewwe et al (Glewwe, Kremer and Moulin, 2009) that four different randomized evaluations have shown that, even in environments where textbooks are relatively scarce, the expansion of availability of textbooks had zero impact on the learning on average. Further investigation of those four studies found that each proposed a different causal mechanism whereby the additional supply of textbooks did not lead to higher learning (for instance in one study in Kenya, the textbooks were too difficult for the typical child, in other, in Sierra Leone the teachers received but did not use the textbooks, in India the textbooks had no incremental impact unless interacted in a treatment arm with performance pay). This just displays that even for something that is known to be an input into learning there are many ways in which program design can cause a failed project and shows that there are potentially many program design interactions even for something as seemingly straightforward as getting textbooks to school children.

Livelihoods There are a variety of programs that transfer assets to poor individuals in an attempt to achieve sustained increases in incomes. In India, the central government has supported states in launching livelihoods programs that create and support women's self-help groups as an instrument to women's empowerment and income gains. A (forthcoming) evaluation of the Jeevika program in Bihar India in the early implementation phase

showed phenomenally large improvements measures of women's autonomy of action in Jeevika villages. However the randomized evaluation of the program scaled statewide (also forthcoming) showed impacts on these same measures that were up to an order of magnitude smaller, often not statistically significant. In this case even the same program design failed to produce the same results when the intensity and integrity of implementation was lessened.

Recently, an experimental evaluation (Banerjee et al., 2015) showed success in an income generation program in five of six countries, suggesting that it is possible to achieve similar results across contexts. In some respects this demonstrates how hard it is to achieve similar results as at all sites the intervention had the same six elements (an asset transfer (e.g. livestock), training on managing the asset, food or cash support, frequent coaching visits, health education/access, a savings account and at all sites and was implemented by exactly the same NGO. This was the scaling across contexts of an approach that had devoted effort and learning the need and how to combined the various dimensions of the design space before moving to an RCT and controlled the implementation afterward.

General Reviews of RCTs

Eva Vivalt (Vivalt, 2016) founded an NGO, AidGrade, that has systematically collected results of over 600 RCTs and has attempted to use that data to ask almost exactly the empirical counter-part of our simulation exercise: how much do RCT results for the same class of intervention differ across studies? This combines all sources of variability across studies: external validity, construct validity, sample variability, and others. She first calculates a standardized impact estimate:

$$SMD = (\mu_{treatment} - \mu_{control})/\sigma_{pooled}$$

She matches interventions and outcomes (e.g. impact of a Conditional Cash Transfer on the enrollment rate) and calculates for each study the intervention impact on outcome. She can then calculate across a class of interventions two measures of the reliability of the impacts: the coefficient of variation and the I^2. The I^2 measure for meta-analysis was

introduced by Higgins and Thompson (Higgins and Thompson, 2002) and "describes the proportion of the total variation in study estimates that is due to heterogeneity." Table 3.7, adapted from Vivalt shows the results for three intervention-outcome pairs with relatively larger numbers of studies and the median across the 51 intervention-outcome pairs with sufficient studies for these estimate to be computed.

The basic finding is that the typical (median) coefficient of variation of impact estimates is 1.77. This implies that if I had a class of RCT studies that showed, on average, a massive impact of 0.5 the one standard deviation across studies would be 0.885 and hence the once standard deviation confidence interval of the impact estimated in the next study would range from 1.38 to *negative* 0.385–that is, if the assumption was the impact of the intervention was to be positive the existing results would be roughly uninformative about the likely result: the plausible range based on the mean and variance of RCT evidence roughly spans the entire plausible range of outcomes as it ranges from very negative to implausibly large. Vivalt reports the typical CV in the medical literature is 0.05 to 0.5. Higgins (Higgins and Thompson, 2002) suggests that for a meta-analysis an I^2 of 0.25 is "low" and 0.75 is "high" and obviously 1.0 is the maximum so the data suggest the heterogeneity in development program RCTs is massive.

Moreover, much of this variation in estimated impacts was "within papers"–that is across different groups or treatment arms (inclusive of differences due to sampling variability) of the same paper estimating the same impact in the same intervention-outcome pair. This is at least suggestive that the traditional interpretation of heterogeneity due to "external validity" caused by "context" is incomplete.

We would argue that these degrees of heterogeneity suggest that the classes of interventions that are often discussed in reviews of evidence about policies/programs/projects exist in a world that is sufficiently "rugged" that they lack construct and external validity and hence are roughly meaningless–it is not clear an "HIV/AID Education project" (or most of the other intervention-outcome pairs Vivalt considers), without further specification of the program design and context, could be the object of meaningful empirical discussion as the

evidence does not predict future outcomes. And yet, Vivalt (2016) reports that one in five studies failed to even report who was responsible for implementation.

We emphasize that this heterogeneity in estimated impacts of RCTs is exactly what was expected as this heterogeneity existed in the non-experimental estimates and Pritchett (Pritchett and Sandefur, 2013) show there is no logically coherent argument that RCT estimates of program impact should produce a lower variance of impact estimates across program design and context than non-experimental estimates. Obviously if the heterogeneity in non-experimental estimates is due to the evaluation of program impacts of different program designs with a rugged fitness function then the heterogeneity is revealing a feature of the world, not of method. This also implies, as (Pritchett and Sandefur, 2015) point out that using rigorous (RCT) evidence from a different context and program design may *increase* the prediction error of true impact over using non-experimental evidence from the same context and design.

3.5.3 Behavioral and Ruggedness

As the advent of the increased use of RCTs in development and the increased popularity of behavioral economics came roughly at the same time, it is worth pointing out that one of the key features of behavioral economics is that it often suggests large impacts across small differences in design where standard theories suggest little or no impact. Perhaps the classic behavioral finding is that whether the default is that a person is signed up for retirement deductions and must opt out or the default is that the person has to actively opt in to retirement deductions has a huge impact on participation and hence the "power of suggestion" Madrian and Shae (Madrian and Shea, 2001) created by the program design feature of the default selection has a much larger role on behavior than standard economic theory would have suggested. Another cited example from RCTs is large changes in demand for health and education services at zero monetary cost pricing (Holla and Kremer, 2009) which is inconsistent with standard economic theory of demand. Zero is considered just another point on the demand curve because of opportunity cost.

Many of these behavioral models imply that the fitness space over program design is rugged, and is rugged in ways that are contextually hard to predict *ex ante*. For example, Gino et al. (Gino, Ayal and Ariely, 2009) demonstrate that Carnegie Mellon students are more likely to cheat at a task when someone observed cheating at the task is wearing a plain shirt and thus assumed to also be a Carnegie Mellon student. But when the cheater is wearing a University of Pittsburgh t-shirt (Carnegie Mellon's cross-town rival), it does not affect cheating by other students. This study suggests that cheating decisions are norm-driven and norm-driven behavior is influenced by identity affinity to the norm violater. But how would one translate that into policy about cheating on property taxes in Pakistan or teacher attendance in Kenya? In another example, Bertrand et al (Bertrand et al., 2010) find that when an advertising mailer includes a photo of someone of the same race the effect of the mailer on loan take-up is nearly twice as large. Precisely as the authors say: "Although it was difficult to predict ex ante which specific advertising features would matter most in this context, the features that do matter have large effects."

Table 3.7: *Variability across RCT Studies for Intervention-Outcome Pairs*

(1) Intervention	(2) Outcome	(3) CV(SMD_i)	(4) Within paper CV	(5) I^2	(6) Number studies
Conditional Cash Transfers	Enrollment Rate	0.83	0.968	1.00	37
HIV/AIDS Education	Use of contraception	3.12	6.97	0.51	10
Micronutrients	Hemoglobin	1.44	0.731	1.00	46
Median (51 intervention/outcome pairs)		**1.77**		**0.99**	**7 (per pair)**

Source: (Vivalt, 2016), Appendix C, Table 12.

3.6 Emerging Learning Mechanisms for Development[12]

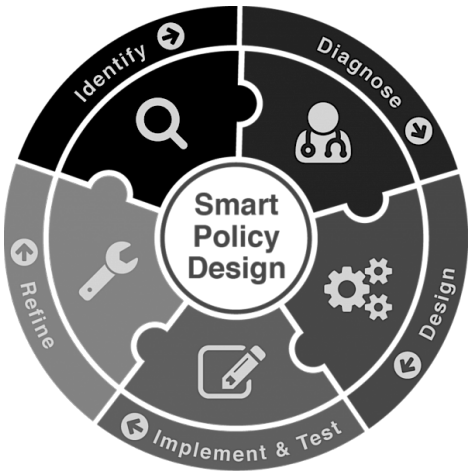
There was a burst of enthusiasm for the use of RCT embedded in "independent impact evaluation" (IIE) as a learning tool for development. A characterization (hopefully not caricature) of this approach is that that many organizations, both NGOs and governments (often financed by donors), were implementing projects and these projects could be paired with "independent" academics/think tanks/consultants who would work with implementers to randomize the project's intervention/treatment so that estimates of project impact on outcomes could achieve "internal validity." The hope was that the proliferation of IIE using RCTs would lead to a body of evidence through, say, "systematic reviews" that would knowledge about "what works" that would produce a superior development practice. This original approach to IIE/RCT as a learning model is now dead, victim of (at least) [13] its own success at creating hundreds of RCTs which have now demonstrated to everyone's satisfaction that the results are too heterogeneous across contexts and program design within "classes" of program and implementers for the type and speed of learning that the traditional IIE/RCT approach provides to be adequate. RCTs have demonstrated that the world of development practice is not a world amenable learning predominantly from a few expensive, slow, rigid impact evaluations of limited treatment arms in projects.

The RCT-IE approach had three significant downsides. One, it was not useful to organizations that are searching for the locally appropriate design. The involvement of actors with an interest in "independent impact evaluation" who want to protect the integrity of the experiment is at odds with those who want to alter the program design in response to real time feedback to make it work better. Second, the granularity of the reality of program design (particularly as informed by behavioral economics) versus attempts to summarize evidence about broad classes of programs such as "micro-credit" or "business training" or

[12] Projects/Policies/Programs

[13] "At least" of this cause as many were convinced of the approach at any time that the approach was never very plausible as it was never grounded in a particularly coherent theory of development realities (Deaton, 2009) or (Rodrik, 2009) or a persuasive positive model of either organizational demand for learning (Pritchett, 2002) or policy maker action in adopting reforms (Pritchett, 2009)

Figure 3.3: *SMART approach to policy design*



"ICT in classrooms" or "job placement" will lead to inappropriate and/or inadequate advice as there is no construct validity of the class of programs. Third, with a lack of construct validity there will be a lack of "external validity" as the results of RCTs will not be predictive of program impact.

Currently, many people are proposing learning mechanisms that do not turn away from randomization as a tool, but bring it into learning strategies embedded with the organization (hence not "independent") and allow for more rapid feedback loops often focused on an earlier part of the causal chain and hence intermediate outputs and outcomes (hence not "impact" evaluation). Just as an example of the changing approach to the use of evaluation and learning:

- MeE (Pritchett, Samji and Hammer, 2013) (Monitoring, experiential learning, impact Evaluation) is a learning approach embedded into a larger strategy for building organizational capability called PDIA (Problem Driven Iterative Adaptation).

- The SMART approach promoted by Evidence for Policy Design (EPoD) (EPoD, 2015) at Center for International Development also emphasizes embedding the feedback loops into the policy design process.

- The group IDinsight (Shah et al., 2013) is making the distinction between KFE (Knowledge-Focused Evaluation) and DFE (Decision-Focused Evaluation) which focuses on using learning to inform decisions about program design during early stage implementation.

- JPAL and IPA are increasingly downplaying the role of "independent" and "impact" evaluation and stressing an engaged learning with partners in a process of program design and its modification.

- The World Bank's research group has supported a "Social Observatory" which engages in both impact evaluation but also in real time feedback on projects particularly on social dimensions of project design, for example women's empowerment (Sanyal, Rao and Majumdar, 2015), deliberative democracy (Besley, Pande and Rao, 2005) and developed methods for this feedback (Bamgartner, Woolcock and Rao, 2010).

3.6.1 Similar Learning Approaches in Other Domains

While the advocacy for the use of RCTs in generating evidence for development often appeals to the analogy of the use of double blind control trials in medicine,[14] it is increasingly recognized in medicine that this as a method of learning is an incomplete (and excessively expensive) approach to evidence and learning.

In a widely cited article, Paul et al (Paul et al., 2010) discuss the decline in productivity and increasing cost of bringing new drugs to market in the traditional paradigm of pharmaceutical R&D which relies on expensive Phase II and Phase III RCT testing. Instead they propose a new approach called "quick win, fast fail" which moves more action into the "proof of concept" phase where costs are lower. By shifting costs from later phases to earlier phases this research can generate more early entities and by providing rapid early feedback can raise the probability of success in the later stages. So, while the RCT is

[14]Ironically, RCTs are often promoted to economists as the Gold Standard even though extremely few economists believe the Gold Standard is the Gold Standard of monetary arrangements.

the Gold Standard for Phase II and III this occupies less of the total research budget and learning process.

Berwick (Berwick, 1998) proposes a process called Plan-Do-Study-Act (PDSA), whereby physicians (or groups of medical-delivery professionals in the form of clinics or hospitals), can plan and implement a change to their process as they see fit, study the success of that change, and either adopt or reject that change upon reviewing the effects. Eppstein et. al (2012) suggest a similar process for learning, Quality Improvement Collaboratives (QIC) also in a medical setting. They propose that a number of agents (hospitals) implement Berwicks PDSA proposal, and share their discovered best practices on an ongoing basis, each adopting recommended practices of the others. Eventually, through several simultaneous PDSA mechanisms, they will converge on an "optimal" program design, whose outputs are a local maximum (adjusting the program design will have reduce the desired output, although it is possible that a more effective project design exists elsewhere in the design space). Eppstein et al (2012) simulate the QIC learning strategy compared with a standard RCT whereby program alterations are made only when they are proven to be significantly more effective and show in simulations that QIC results in a more effective program than a typical RCT in nearly all cases. RCT is superior only in the highly idealized scenario that the design space is non-rugged and the number of observations in each iteration is quite high. So while the RCT is the Gold Standard for drug approvals it is not necessarily the best learning stategy for improving medical practices in complex organizational settings.

The concept of a Realist Evaluation has appeared largely in the public health literature (Pawson et al., 2005). A Realist Evaluation identifies three key variables: the Context (C) in which a program is implemented, and the Mechanism (M) through which the program has the desired Outcome (O). As Pawson et al state, the question in a Realist Evaluation becomes *What is it about this program that works for whom in what circumstances?* thus limiting the breadth of the question and identifying the relationship between the implementation and the outcome. A realist evaluation of a leadership development program in Ghana (Kwamie, van Dijk and Agyepong, 2014) relied on an explanatory case study of the program.

In addition to H1, the hypothesis they seek to prove or disprove, the authors detail H0: a parallel, alternative hypothesis, which would exist should the proposed hypothesis be rejected, and looked for both characteristics of H1 and of H0 in their analysis. Their research uses a combination of collected data, observation, document review, and semi-structured interviews. Upon finding several examples that support H0, they rejected their original hypothesis.

Through the Lean Startup Methodology, the ongoing optimization process has been organized into a popular learning methodology in the startup community. Eric Reiss (Reiss, 2011) describes the methodology as ”Ideas - Code - Data” where a concept is determined, implemented, tested, and then improved upon to start the circle all over again. Reiss argues that the benefits of learning about your product, how it is used, and whether it meets the needs of your target customers outweigh the costs of going to market too quickly.

The Lean Startup Methodology, and similar processes recommend that entrepreneurs follow a specific model of carefully identifying the problem they aim to solve and characteristics of that market, and then design small experiments to determine whether their hypothesis that their product will solve that problem is correct. At the earliest stage, the small experiment could be speaking with people on the street. In later stages, it might be releasing a beta version of the product and determining ahead of time the number of users after 24 hours that would demonstrate that the product is on the right track. It is notable in that entrepreneurs are advised to determine their decision-making rule before running their study, and adjusting some aspect of the product design should the study not obtain the desired results. There is no room for explaining away the results and continuing on the same path. The concept of applying the Lean Startup Methodology to social programs has already gained traction. Acumen+ offers a course called “Lean Startup Principles for Social Impact” (website, 8/20/2015), and Lean Impact for Social Good organizes summits and learning opportunities about applying the Lean Startup Principles to social programs.

3.7 Conclusion

If the world in which development programs/projects/policies are conceived of, designed and implemented is one that (a) has high dimensional design spaces, (b) has rugged fitness functions over those design spaces and (c) has a fitness function that is contextual (in many and perhaps unknown ways), then the standard approach of "theory, design, implement with an impact evaluation" is unlikely to be successful and is certainly not the optimal approach. Rather, the process has to involve a (potentially extended) period in which learning comes from rapid feedback on movements across the design space to reach effective (if not optimal) program designs. The first two incarnations of the *randomista* movement ("independent impact evaluation" and "do-it-ourselves" RCTs) are being rapidly replaced with a variety of approaches to development that are more sophisticated about organizational learning.

Bibliography

Alatas, Vivi, Abhijit Banerjee, Rema Hanna, Benjamin A Olken, and Julia Tobias. 2010. "Targeting the Poor: Evidence from a Field Experiment in Indonesia." National Bureau of Economic Research.

Angelucci, Manuela, and Giacomo De Giorgi. 2009. "Indirect Effects of an Aid Program: How Do Cash Transfers Affect Ineligibles' Consumption?" *The American Economic Review*, 486–508.

Angelucci, Manuela, and Orazio Attanasio. 2009. "Oportunidades: Program Effect on Consumption, Low Participation, and Methodological Issues." *Economic development and cultural change*, 57(3): 479–506.

Attanasio, Orazio, and Valérie Lechene. 2010. "Conditional cash transfers, women and the demand for food." IFS working papers.

Bamgartner, Michael, Michael Woolcock, and Vijayendra Rao. 2010. "Using Mixed Methods In Monitoring And Evaluation : Experiences From International Development." *World Bank Policy Research Working Paper No. 5245*.

Banerjee, Abhijit, Esther Duflo, Nathanael Goldberg, Dean Karlan, Robert Osei, William Parient, Jeremy Shapiro, Bram Thuysbaert, and Christopher Udry. 2015. "A Multifaceted Program Causes Lasting Progress for the very Poor: Evidence from Six Countries." *Science*, 348.

Bertrand, Marianne, Dean Karlan, Sendhil Mullainathan, Eldar Shafir, and Jonathan Zinman. 2010. "What's Advertising Content Worth? Evidence from a Consumer Credit Marketing Field Experiment." *The Quarterly Journal of Economics*, 125(1): 263–306.

Berwick, Donald M. 1998. "Developing and testing changes in delivery of care." *Annals of Internal Medicine*, 128(8): 651–656.

Besley, T, and R Kanbur. 1993. "Principles of targeting'University of Warwick Development Economics Research Discussion Papers No. 85."

Besley, Timothy, Rohini Pande, and Vijayendra Rao. 2005. "Participatory Democracy in Action: Survey Evidence from South India." *Journal of the European Economic Association*, 3(2-3): 648–657.

Bloom, Nicholas, Aprajit Mahajan, David McKenzie, and John Roberts. 2010. "Why do Firms in Developing Countries have Low Productivity?" *The American Economic Review*, 619–623.

Bourguignon, François, Francisco HG Ferreira, and Phillippe G Leite. 2002. "Ex-ante Evaluation of Conditional Cash Transfer Programs: the Case of Bolsa Escola." *World Bank Policy Research Working Paper*, , (2916).

Bradley-Geist, Jill C, and Ronald S Landis. 2012. "Homogeneity of Personality in Occupations and Organizations: A Comparison of Alternative Statistical Tests." *Journal of Business and Psychology*, 27(2): 149–159.

Breaugh, James A, and Rebecca B Mann. 1984. "Recruiting Source Effects: A test of Two Alternative Explanations." *Journal of Occupational Psychology*, 57(4): 261–267.

Camacho, Adriana, and Emily Conover. 2011. "Manipulation of Social Program Eligibility." *American Economic Journal: Economic Policy*, 41–65.

Chetty, Raj. 2012. "Bounds on Elasticities with Optimization Frictions: A Synthesis of Micro and Macro Evidence on Labor Supply." *Econometrica*, 80(3): 969–1018.

Copestake, James, Monica Guillen-Royo, Wan-Jung Chou, Tim Hinks, and Jackeline Velazco. 2009. "The Relationship between Economic and Subjective Wellbeing Indicators in Peru." *Applied Research in Quality of Life*, 4(2): 155–177.

Das, Jishnu, Stefan Dercon, James Habyarimana, Pramila Krishnan, Karthik Muralidharan, and Venkatesh Sundararaman. 2011. "School Inputs, Household Substitution, and Test Scores." , (16830).

Deaton, Angus S. 2009. "Instruments of Development: Randomization in the Tropics, and the Search for the Elusive Keys to Economic Development." *National Bureau of Economic Research*, January(14690).

De Janvry, Alain, Frederico Finan, Elisabeth Sadoulet, and Renos Vakis. 2006. "Can Conditional Cash Transfer Programs Serve as Safety Nets in Keeping Children at School and from Working when Exposed to Shocks?" *Journal of Development Economics*, 79(2): 349–373.

EPoD. 2015.

Evans, David, and Anna Popova. 2015. "What Really Works to Improve Learning in Developing Countries? an Analysis of Divergent Findings in Systematic Reviews." *An Analysis of Divergent Findings in Systematic Reviews (February 26, 2015). World Bank Policy Research Working Paper*, , (7203).

Fafchamps, Marcel, and Alexander Moradi. 2009. "Referral and Job Performance: Evidence from the Ghana Colonial Army." *CEPR Discussion Paper No. DP7408*.

Gino, Francesca, Shahar Ayal, and Dan Ariely. 2009. "Contagion and differentiation in unethical behavior the effect of one bad apple on the barrel." *Psychological Science*, 20(3): 393–398.

Glewwe, Paul, Michael Kremer, and Sylvie Moulin. 2009. "Many Children Left Behind? Textbooks and Test Scores in Kenya." *American Economic Journal: Applied Economics*, 1(1): 112–35.

Guillen-Royo, Monica. 2008. "Consumption and Subjective Wellbeing: Exploring Basic Needs, Social Comparison, Social Integration and Hedonism in Peru." *Social Indicators Research*, 89(3): 535–555.

Heckman, James J. 2006. "Skill formation and the economics of investing in disadvantaged children." *Science*, 312(5782): 1900–1902.

Heckman, James J, Jora Stixrud, and Sergio Urzua. 2006. "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior." *Journal of Labor Economics*, 24(3): 411–482.

Higgins, Julian P. T., and Simon G. Thompson. 2002. "Quantifying Heterogeneity in Meta-Analysis." *Statistics in Medicine*, 21: 1539–1558.

Holla, Alaka, and Michael Kremer. 2009. "Pricing and Access: Lessons from Randomized Evaluations in Education and Health." *Center for Global Development Working Paper*, 158(January).

Holland, J. 1985. *Making Vocational Choices.* Englewood Cliffs, NJ: Prentice Hall. Kanter, RM.

Jensen, Robert. 2010. "The (Perceived) Returns to Education and the Demand for Schooling." *The Quarterly Journal of Economics*, 125(2): 515–548.

Kauffman, Stuart, and Simon Levin. 1987. "Towards a General Theory of Adaptive Walks on Rugged Landscapes." *Journal of theoretical Biology*, 128(1): 11–45.

Kinsler, Josh, and Ronni Pavan. 2012. "The Specificity of General Human Capital: Evidence from College Major Choice." working paper.

Kwamie, Aku, Han van Dijk, and Irene Akua Agyepong. 2014. "Advancing the Application of Systems Thinking in Health: Realist Evaluation of the Leadership Development Programme for District Manager Decision-making in Ghana." *Health Res Policy Syst*, 12(29): 10–1186.

Lee, David S, and David Card. 2008. "Regression Discontinuity Inference with Specification Error." *Journal of Econometrics*, 142(2): 655–674.

Levy, Dan, and Jim Ohls. 2010. "Evaluation of Jamaica's PATH conditional cash transfer programme." *Journal of Development Effectiveness*, 2(4): 421–441.

Loury, Linda Datcher. 2006. "Some contacts are more equal than others: Informal networks, job tenure, and wages." *Journal of Labor Economics*, 24(2): 299–318.

Madrian, B, and D Shea. 2001. "The power of suggestion." *Quarterly Journal of Economics*, 18–116.

Maluccio, John A., Michelle Adato, Rafael Flores, and Terry Roopnaraine. 2005. "Nicaragua: Breaking the Cycle of Poverty." International Food Policy Research Institute.

Mbiti, Isaac, Karthik Muralidharan, and Youdi Schipper. 2015. "Inputs, Incentives, and Complementarities in Primary Education: Experimental Evidence from Tanzania." *in process*.

McEwan, Patrick. 2015. "Improving Learning in Primary Schools of Developing Countries." *Review of Educational Research*, 85(3): 353–394.

Ministerio de Trabajo del Peru Dataset. 2012.

Muralidharan, Karthik. 2015. "Comments at RISE Meeting."

Nichols, Albert L, and Richard J Zeckhauser. 1982. "Targeting Transfers through Restrictions on Recipients." *The American Economic Review*, 372–377.

Ortiz, Viviola Gómez. 2010. "Assessment of Psychosocial Stressors at Work: Psychometric Properties of the Spanish Version of the ERI (Effort-Reward Imbalance) Questionnaire in Colombian Workers." *Revista de Psicología del Trabajo y de las Organizaciones*, 26(2): 147–156.

Pallais, Amanda. 2013. "Small Differences that Matter: Mistakes in Applying to College." National Bureau of Economic Research.

Panatik, Siti Aisyah Binti, Azizah Rajab, Roziana Shaari, Maisarah Mohamed Saat, Shahrollah Abdul Wahab, and Nurul Farhana Mohd Noordin. 2012. "Psychosocial Work Condition and Work Attitudes: Testing of the Effort-Reward Imbalance Model in Malaysia." *Procedia-Social and Behavioral Sciences*, 40: 591–595.

Paul, Steven M., Daniel S. Mytelka, Christopher T. Dunwiddie, Charles C. Persinger, Bernard H. Munos, Stacy R. Lindborg, and Aaron L. Schacht. 2010. "How to improve R&D Productivity: the Pharmaceutical Industry's Grand Challenge." *Nature Reviews Drug Discovery*, 9: 203–214.

Pawson, Ray, Trisha Greenhalgh, Gill Harvey, and Kieran Walshe. 2005. "Realist Review–a New Method of Systematic Review Designed for Complex Policy Interventions." *Journal of Health Services Research & Policy*, 10(suppl 1): 21–34.

Phillips, Leonard W. 1968. "Occupational Choice and Vocational Interests." *The Journal of Educational Research*, 61(8): 355–359.

Pritchett, Lant. 2001. "Where has All the Education Gone?" *The World Bank Economic Review*, 15(3): 367–391.

Pritchett, Lant. 2002. "It Pays to be Ignorant: A Simple Political Economy of Rigorous Program Evaluation." *The Journal of Policy Reform*, 5(4): 251–269.

Pritchett, Lant. 2009. "The Policy Irrelevance of the Economics of Education: Is Normative as Positive Just Useless, or Worse?" In *What Works in Development?: Thinking Big and Thinking Small.* , ed. Jessica Cohen and William Easterly. Brookings Institution Press.

Pritchett, Lant, and Justin Sandefur. 2013. "Context Matters for Size: Why External Validity Claims and Development Practice Don't Mix." *Center for Global Development Working Paper*, 336(August).

Pritchett, Lant, and Justin Sandefur. 2015. "Learning from Experiments When Context Matters." *American Economic Review*, 105(5): 471–475.

Pritchett, Lant, Salimah Samji, and Jeffrey S Hammer. 2013. "It's all about MeE: Using Structured Experiential Learning ('e') to crawl the design space." *Center for Global Development Working Paper*, , (322).

Reiss, Eric. 2011. *The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses.* Crown Business.

Rodrik, Dani. 2009. "The New Development Economics: We Shall Experiment, but How Shall We Learn?" In *What Works in Development?: Thinking Big and Thinking Small.* , ed. Jessica Cohen and William Easterly. Brookings Institution Press.

Roy, Andrew Donald. 1951. "Some Thoughts on the Distribution of Earnings." *Oxford Economic Papers*, 3(2): 135–146.

Rubinstein, Yona, James J Heckman, et al. 2001. "The Importance of Noncognitive Skills: Lessons from the GED Testing Program." *American Economic Review*, 91(2): 145–149.

Sabarwal, Shwetlena, David K. Evans, and Anastasia Marshak. 2014. "The Permanent Input Hypothesis : the Case of Textbooks and (no) Student Learning in Sierra Leone." *World Bank Policy Research Working Paper*, , (WPS 7021).

Saez, Emmanuel. 2010. "Do Taxpayers Bunch at Kink Points?" *American Economic Journal: Economic Policy*, 180–212.

Sanyal, Paromita, Vijayendra Rao, and Shruti Majumdar. 2015. "Recasting Culture to Undo Gender: A Sociological Analysis of Jeevika in Rural Bihar, India." *World Bank Policy Research Working Paper No. 7411*.

Satterwhite, Robert C, John W Fleenor, Phillip W Braddy, Jack Feldman, and Linda Hoopes. 2009. "A case for homogeneity of personality at the occupational level." *International Journal of Selection and Assessment*, 17(2): 154–164.

Schultz, T Paul. 2004. "School Subsidies for the Poor: Evaluating the Mexican Progresa Poverty Program." *Journal of development Economics*, 74(1): 199–250.

Schwarzweller, Harry K. 1960. "Values and Occupational Choice." *Social Forces*, 39(2): 126–135.

Shah, Neil Buddy, Paul Wang, Fraker Andrew, and Daniel Gastfriend. 2013. "Evaluations with Impact Decision-focused Impact Evaluation as a Practical Policymaking Tool." International Initiative for Impact Evaluation Working Paper 25.

Siegrist, Johannes, Jian Li, and Diego Montano. 2014. "The Psychometric properties of the Effort-Reward Imbalance Questionnair." Department of Medical Sociology, Faculty of Medicine, Duesseldorf University, German.

Simpson, Richard L, and Ida Harper Simpson. 1960. "Values, Personal Influence, and Occupational Choice." *Social Forces*, 39: 116.

Skoufias, Emmanuel, Susan W Parker, Jere R Behrman, and Carola Pessino. 2001. "Conditional Cash Transfers and their Impact on Child Work and Schooling: Evidence from the Progresa Program in Mexico [with comments]." *Economia*, 45–96.

Slemrod, Joel. 2007. "Cheating Ourselves: The Economics of Tax Evasion." *The Journal of Economic Perspectives*, 25–48.

Spence, Michael. 1973. "Job Market Signaling." *The Quarterly Journal of Economics*, 355–374.

Sullivan, Paul. 2010. "A Dynamic Analysis Of Educational Attainment, Occupational Choices, And Job Search." *International Economic Review*, 51(1): 289–317.

Tybout, James R. 2000. "Manufacturing Firms in Developing Countries: How Well do they Do, and Why?" *Journal of Economic literature*, 11–44.

Van der Klaauw, Wilbert. 2012. "On the Use of Expectations Data in Estimating Structural Dynamic Choice Models." *Journal of Labor Economics*, 30(3): 521–554.

Vivalt, Eva. 2016. "How Much can we Generalize from Impact Evaluation Results?"

Appendix A

Appendix to Chapter 1

A.0.1 Bios of Contributors to the Test Advisory Committee

The following people supported the development of the questions in this evaluation through helping write questions, supporting internal validation analysis, or reviewing questions for comprehension and cultural appropriateness.

- **Kevin Joldersma:** PhD in Measurement and Quantiative Methods from Michigan State University, Masters in Spanish Language from University of Illinois, Urbana Champaign. Has worked as a psychometrician at Second Language Testing Inc., Amplify Learning, and Questar Assessment.
- **Adriana González Ríos:** Masters in Education from the Harvard School of Education. Co-Founder of Scholastica, an educational consulting firm that prepares Mexican applicants for US educational testing.
- **Ed Gaible:** Founder and Principal at Natoma Group Consulting, which provides consulting services for technology in education projects for development.
- **Mary Burns:** Masters in Education from the Harvard School of Education and in Latin American Studies and Urban Planning from the University of Texas, Austin. Professional development specialist and researcher at the Education Development Center.

A.0.2 Supplementary Tables to Chapter 1

Table A.1: *Preferences for Work and Number of Family Members with a Formal Job*

Preferences for work and number of family members with a formal job

	(1)	(2)	(3)	(4)
	Family	Income	Achievement	Quality
	b/se	b/se	b/se	b/se
Business Administration	0.146	0.0860	0.894**	-1.283***
	(0.537)	(0.323)	(0.184)	(0.141)
Engineering	0.262	0.117	-0.0538	-0.821
	(0.265)	(0.488)	(0.214)	(0.527)
Num HH Members Working	0.0293	0.442	0.145	-0.468*
	(0.288)	(0.235)	(0.102)	(0.167)
Business * Num HH Members Working	-0.00750	-0.155	-0.380	0.467
	(0.401)	(0.352)	(0.193)	(0.238)
Engineering * Num HH Members Working	-0.136	-0.106	0.168	0.164
	(0.186)	(0.317)	(0.135)	(0.363)
Constant	2.953	2.693*	3.461**	5.359***
	(1.266)	(0.943)	(0.762)	(0.445)
r2	0.0139	0.0271	0.0314	0.0689
N	677	677	677	677

Control for University Location, Age, Age^2, Gender; Standard Errors clustered at location level—
* p <0.10, ** p <0.05, *** p <0.01

Appendix B

Appendix to Chapter 2

B.1 A Model of Asset Aquisition

Suppose that income is a linear function of wages, w and Progresa transfers T, and wages have a mean $\bar{w}$ plus a time-specific shock with mean 0.

$$y_t = w_t + P_t \tag{B.1}$$

$$w_t = \bar{w} + \xi_t \tag{B.2}$$

Asset ownership in the current period is a function of income and asset ownership in the previous period. In period t, households purchase assets based on their income in period $t-1$, and asset ownership in period $t-1$. Their income in period t is not revealed until the end of the period. [1]

$$A_t = g(y_{t-1}, A_{t-1}(y_{t-2})) \tag{B.3}$$

Where $g'_y > 0$ and $g'_A > 0$.[2] There is some $y = \bar{y}$ for which $g''_{y \leq \bar{y}} > 0$ and $g''_{\bar{y} \leq y} < 0$. This is consistent with the S-curve.

[1] To the extent that current income is a proxy for future income, this equation applies, even in fully-functioning credit markets.

[2] Households are more likely to own an asset if they owned it in the previous period, or if they have higher income.

The relationship between current and previous period asset ownership can be extended back to period 0:

$$A_t = h((t-1)\bar{w}\sum_{s=0}^{t-1}((1+\xi_s)+\frac{P_s}{w})) \tag{B.4}$$

This highlights another benefit of PMT scores: they are a more accurate reflection of wealth (lifetime earnings to date, and expected lifetime earnings) than current income in populations with variable incomes, such as Mexico's rural poor.

When I match households by baseline PMT score, I look at households that were identical at baseline (in this analysis, the equivalent of period $t-2$), but their likelihood of aquiring an asset in period $t-1$, when Progresa transfers were given, differs from one another. Thus, the estimate of the "incentive effect" (distortions resulting from the implicit tax) is not isolated from the "wealth effect" (purchases made due to the additional income from Progresa):

$$\begin{aligned} E[Y|D=1] - E[Y|D=0] = & \\ &= [h((\bar{w}+\xi_s+T_s)+(t-2)\bar{w}\sum_{s=0}^{t-2}(\xi_s))] \\ &- [h((\bar{w}+\xi_s)+(t-2)\bar{w}\sum_{s=0}^{t-2}(\xi_s))] \end{aligned} \tag{B.5}$$

B.2 Household Spending Items

Table B.1: *Spending Items included in Total Household Spending*

Food	**Other**
Onion	School Transportation
Potato	Other Transportation
Carrot	Tobacco
Greens	Hygiene Products
Orange	Gas
Banana	Electricity
Apple	Kitchen Goods
Lemon	Other Household Goods
Cactus	Linens
Other Vegetables	Men's Clothing
Sauce	Women's Clothing
Soda	Boys' Clothing
Alcohol	Girls' Clothing
Coffee	Men' s Shoes
Sugar	Women's Shoes
Oil	Boys' Shoes
Chicken	Girls' Shoes
Beef	Ceremonies
Goat	School PTA
Fish	
Eggs	
Milk	
Cheese	
Pork Fat	
Other Animals	
Tortilla	
Cornmeal	
White Bread	
Sweet Bread	
Fresh Bread	
Flour	
Pasta	
Rice	
Crackers	
Beans	
Cereal	
Other Grains	

B.3 Probabilistic Response

In this case, households will optimize over probabilities of whether the house would receive the PMT score. The transfer, T is not assured, but can only be expected $\pi(a_i)$ portion of the time. Households deciding whether to adjust spending on assets will choose the optimal of the following three options:

$$\begin{aligned}\max_{a^h_{i,t}, a^{nh}_{i,t}} \{ & \pi(a^{nh}_{i,t}) \left((a^h_{i,t} + a^{nh}_{i,t})^{\beta_i} (y + a_{i,t-1} + T - h(a^h_{i,t}))^{1-\beta_i} \right) \\ & + (1 - \pi(a^{nh}_{i,t})) \left((a^h_{i,t} + a^{nh}_{i,t})^{\beta_i} (y + a_{i,t-1} - h(a^h_{i,t})^{1-\beta_i}) \right) \}\end{aligned} \tag{B.6}$$